Mining and learning from spatio-temporal data: examples and perspectives

Dino Ienco

HABILITATION Á DIRIGER DES RECHERCHES

UNIV. MONTPELLIER, ECOLE DOCTORALE I2S, INFORMATIQUE

Mining and Learning from Spatio-Temporal Data: Examples and Perspectives

With a major emphasis on Remote Sensing Satellite Images

Dino Ienco

IRSTEA
UMR TETIS
Montpellier, France

November 9, 2016

JURY

Dennis SHASHA	Professeur, NYU CS Dept., USA	Rapporteur
Dino PEDRESCHI	Professeur, Universitá di Pisa, Italy	Rapporteur
Patrick GALLINARI	Professeur, Univ. Pierre et Marie Curie, LIP6, Paris	Rapporteur
Jean-François BOULICAUT	Professeur, Insa, Lyon	Examinateur
Bruno CREMILLEUX	Professeur, Univ. de Caen Normandie, Caen	Examinateur
Maguelonne TEISSEIRE	Directeur de Recherche, IRSTEA, Montpellier	Examinateur

Acknowledgements

First of all, I would thank my colleagues at the UMR TETIS and at LIRMM for their fruitful discussion and exchanges. I thank the members of the Advanse team for their positive influence on my researches and on the person who I am now. During these years, I had also the possibility and the privilege to collaborate with many researchers outside France. These collaborations teach me a lot and help me to grow both professionally and personally. I would also thank the members of the jury to have accepted to review my HDR. Last but not least, I thank all my family for their education, support and love they gave me.

Contents

Chapter 1

Introduction

1.1 Abstract

My research activities are related to the fields of Data Mining, Data Base and Machine Learning. The main goal of my work is the development of new techniques and algorithms to manage and analyze large amounts of heterogeneous data with a major emphasis on data involving spatial and temporal characteristics (i.e. satellite images, environmental data, sensor data, etc...).

1.2 Context

During my PhD (2007-2010) I focused my attention on the study of methods and techniques to manage and mine large amounts of information to automatically extract hierarchical representations of the data. More precisely, I designed, developed and implemented clustering techniques to manage and mine textual and categorical data. The results of my researches were: i) methods to evaluate the distance between the data described by categorical variables and ii) techniques to extract hierarchical representations from textual data [44]. In the same period, I also investigated the field of associative classification proposing new classification methods based on local features extracted through itemset mining approaches [66].

During my PhD training period, I also had the opportunity to perform an internship at the Yahoo Research Laboratory in Barcelona for a period of three months where I worked on the analysis of information propagation on social network data [13].

From February 2011 to September 2011, I turned my attention to the analysis of spatio-temporal data. Indeed, data with a strong geographic component are now becoming more widespread and they are a real source of information that poses new challenges to data mining approaches. To improve my skills in this field I applied for a post-doctoral fellowship at Cemagref Institute in Montpellier (now IRSTEA). The topic of the post-doc fellowship was related to the analysis of spatio-temporal data and classification of remote sensing satellite images. The scholarship allowed me to develop skills and propose new innovative approaches to deal with classification issue in the context of time series of remote sensing satellite images. During this period, I have also started a collaboration with the LIRMM lab that is formally described by my association to the TATOO team before and, now, to the recently created ADVANSE team.

Activities after 2011

In September 2011, I was recruited at IRSTEA on the subject of data analysis and knowledge extraction from spatial and temporal data. In particular, my main activities are devoted to study, formalize and develop new data mining and machine learning techniques for spatio-temporal data. From a methodological point of view, I investigated new methods (supervised/unsupervised classification and pattern mining) especially tailored for such kind of data. Most of the techniques I developed model the data through a network (or graph) structure to explicitly represent the spatial interactions among the data. As a privileged field of application, the developed methods are employed for the analysis of remote sensing images where both spatial and temporal components play a crucial role in the knowledge extraction process.

Main Research Activities

In 2012 I co-supervised the PhD thesis of Phan Nhat Hai on moving object data mining with Dr. Maguelonne Teisseire (IRSTEA) and Prof. Pascal Poncelet (LIRMM). The objectives of this thesis were to find new approaches and methods to efficiently mine different kinds of trajectory patterns at the same time [35, 37]. This thesis also contributed to develop new strategies to summarize interesting and useful patterns from the extracted results [36]. Reducing the output size of pattern mining algorithms is an important issue since, sometimes, the size of outputted patterns can be bigger than the original data size. Supplying a method to filter out the most interesting trajectories can thus increase the understandability and the usefulness of the extracted knowledge. The thesis was defended in October 2013 and Dr. Phan Nhat Hai is currently post-doc at the University of Oregon (USA).

In the same year, between February and April 2013, I spent three months as a visiting researcher at the University of Waikato, New Zealand, where I collaborated with Prof. Bernhard Pfahringer to new data mining algorithms for data streams. The main challenge in data stream mining is the severe computational constraint imposed by the high speed at which data arrives. This data velocity can drastically affect the whole mining process [49, 48].

Still in 2013 I started a collaboration with Dr. Andrea Tagarelli (University of Calabria, Italy). He had previously been the reviewer of my PhD thesis. In conjunction with Dr. Andrea Tagarelli, we supervised the PhD thesis of Salvatore Romeo on the use of matrix and tensor decomposition methods for document clustering [79, 78]. Dr. Salvatore Romeo defended his thesis in April 2015 and, since October 2015, he is post-doc researcher at QCRI (Qatar Computer Research Institute, Doha, Qatar) working on Information Retrieval and Natural Language Processing. With Dr. Tagarelli, we are currently collaborating on the development of new data mining approaches for the study of complex and heterogeneous data with a major emphasis on multilingual corpora.

Between 2013 and 2015, I co-supervised Dr. Fabio Güttler, post-doc researcher at UMR TETIS, with Dr. Maguelonne Teisseire and Prof. Pascal Poncelet on the topic of Change Detection analysis of time series of satellite images. The fellowship had a duration of 18 months and it was funded

by the EQUIPEX GEOSUD[1]. This supervision gave me the opportunity to open new research activities in the field of spatio-temporal data analysis [31, 32]. In this work, we proposed new approaches (validated by experts) to better characterize evolution and/or changes in time series of satellite images by proposing a compact and efficient description to depict how natural phenomena evolve over time.. The fellowship ended in March 2015 and Dr. Fabio Güttler is now post-doc at the University of Strasbourg.

In 2014, I started the co-supervision of the PhD thesis of Vijay Ingalalli with Prof. Pascal Poncelet. This thesis is founded half by the Labex Numev and half by a grant that I have obtained from IRSTEA. This thesis focuses on the development of new graph management and mining techniques with applications on remote sensing data. More in detail, the work of Vijay Ingalalli focuses on mulitgraph data. A multigraph is a graph where pairs of nodes can be linked to each other via different types of edges. The multigraph representation is particularly useful to describe and reason about entities that interact with each other through different dimensions. In the case of spatio-temporal data, graph (or multigraph) models are well suited to describe spatial correlations among data. Firstly, we started to work on efficient algorithms to deal with the iso/homomorphism problems in multigraph structures [90, 91]. Now, we are investigating new efficient methods to extract frequent patterns from multigraph data.

In June 2014, I visited the CNR Institute (Milano, Italy) hosted by Dr. Gloria Bordogna with whom I started a collaboration on spatio-temporal data clustering [16, 7]. In the autumn of the same year I was invited by Prof. André Carvalho (USP Sao Carlos, Brazil) as an invited lecturer at the Brazilian doctoral school on Machine Learning and Data Analysis[2]. The topic of the lecture was about advanced Machine Learning approaches (Active Learning).

In autumn 2015 I started the co-supervision of two new PhD thesis on the analysis of remote sensing images.

The thesis of Lynda Khiali, co-supervised with Dr. Maguelonne Teisseire, is founded by an AVERROES scholarship (Algerian government). The main goal is to investigate the analysis of long time series of satellite images. The work is devoted to analyze the image archive Spot World Heritage provided by the CNES institute. This archive contains time series of satellite images that span over a period of more than twenty years. The temporal richness of this source of data can pave the way to better understand complex phenomena such as climate change behaviors, development of wetlands, ecological changes in the study area, etc...

The thesis of Lionel Pibrel, co-supervised with Dr. Marc Chaumont (LIRMM) and Dr. Gerard Subsol (LIRMM) is founded by an ANR CIFRE fellowship in collaboration with the company Berger Levrault. The work focuses on the use of deep learning techniques on heterogeneous remote sensing data for the object detection task in urban areas. Standard classification algorithms need a training phase before being employed on new unseen examples (test data). The goal of this thesis is to develop new deep learning architectures to manage heterogeneous data coming from different sensors (optical sensor, satellite images, drone images, radar images, etc.)

[1] http://ids.equipex-geosud.fr/ (accessed April 5th, 2016)

[2] http://www.amda.icmc.usp.br/mlkdd2014/

in a setting where training and test data cannot fit the same format and, commonly, the test set is poorer (in terms of information) than the training set.

1.3 Thesis Organization

I choose to present the research activities I did in the past five years in the TETIS and LIRMM laboratories by means of a collection of publications. Each paper covers one of the different aspects I have dealt with in the analysis of spatio-temporal data. In addition to these representative publications, I develop three chapters to explain from where I come from, what I have done and what I would like to do in the near future.

More in detail, the first chapter has quickly introduced my career, the research activities I developed during my PhD and the researches I investigated in the last five year once I have joined my current team in Montpellier.

In the second chapter of this manuscript I draw, with a particular emphasis on scientific contributions, a picture about my activities in the field of data mining and machine learning. The second chapter is especially dedicated to my activities in the field of spatio-temporal data analysis, but it also supplies an overview of my skills and background in the general field of data science. It also points out the national/international collaborations I grew in the last years. All these collaborations, I think, heavily highlight that research is not a solitary process but needs exchanges and partnerships in order to benefit from different points of view with the goal of creating new knowledge.

In the third and last chapter, I draw the conclusion of my manuscript with a list of perspectives directly related to my research but, also, to the field of spatio-temporal data analysis. This chapter summarizes a part of my ideas about possible future researches and, possible interesting trends that are currently investigated by the research community. In some sense, it paves the way to future subjects tightly related to my on-going projects that I will be glad to investigate.

Chapter 2

Contributions in Spatio-Temporal Analysis

2.1 Introduction

As the world becomes interconnected, spatio-temporal data are more ubiquitous and are getting more and more attention. Moving object (e.g., taxi, bird) trajectories recorded by GPS devices, social event (e.g., microblogs, crime) with location tag and time stamps, and environment monitoring (e.g., remote sensing images) are typical spatio-temporal data that we meet every day[1]. These emerging spatio-temporal data also bring new challenges and opportunities to data mining and machine learning researchers. This chapter introduces some of the contributions we realized in the past years regarding spatio-temporal data analysis. It is organized in five parts: the first two sections are more related to works on pattern mining and pattern extraction we developed during the researches conducted in Montpellier. The first section is related to the analysis of Moving Object data realized during the PhD Thesis of Dr. Phan Nhat Hai while the second one introduces my recent experiences related to manage spatial interaction through graph-based approaches.

On the other hand, the third and fourth sections are more focused on classification techniques (supervised and semi-supervised) we conceived in partnerships with other colleagues. The third section involves the work on data stream analysis we developed in collaboration with Prof. Bernhard Pfahringer during my visit at the University of Waikato, New Zealand. The fourth section describes some applications of my research in the field of remote sensing image analysis.

Finally, the fifth section gives a quick overview of my research collaborations in the general field of data mining and machine learning.

As a general guideline to read this chapter, the contributions we proposed in the different domains are highlighted in bold face.

2.2 Moving Object Data: Efficiently extract New, Useful and Non Redundant Trajectory Patterns

Techniques able to summarize the behavior of groups of objects moving together are getting more and more attention due to the rapid development

[1] http://researcher.watson.ibm.com/researcher/view_group.php?id=4152

of positioning technologies (smartphones, GPS tracking systems, location-based services, etc..) that constantly record and store position information.

Analyzing data generated from these systems can be useful for:

+ Understanding animal migrations to support public policies in order to preserve biodiversity;
+ Studying and monitor traffic on road networks to better design future transportation systems;
+ Detecting suspicious or anomaly movement patterns behaviors.

This is why during the PhD Thesis of Phan Nhat Hai, co-supervised with Prof. Pascal Poncelet and Dr. Maguelonne Teisseire we studied, conceived and developed new data mining algorithm to extract and summarize collections of trajectory patterns from moving object dataset.

2.2.1 Mine different patterns under a unique approach

A lot of research was done to analyze such datasets with the goal to extract meaningful patterns [38] and, at the same time, many algorithms have been proposed such as $CuTS^*$ [51] (convoy mining), $ObjectGrowth$ [57] (closed swarm mining), $VG\text{-}Growth$ [96] (group pattern mining), etc... Each of the different proposed methods has its own characteristics and it only extracts one type of pattern. This means that if we want to compare different moving object patterns extracted from the same dataset, we need to re-implement each of the different methods and then run each of the algorithm independently.

This issue was addressed by the GET_MOVE **approach [34].** This framework represents the moving object database by a *cluster matrix* in which a row is an object and a column is a cluster of objects at a certain time stamps. The matrix representation is successively employed to extract frequent closed itemsets from which movement patterns can be mined.

2.2.2 Flexible Moving Object Patterns

One issue shared by all the previous proposed moving object pattern methods regards the way parameters are set. Defining a unique strict threshold, i.e. the maximum time gap between pair of object clusters, without some degree of flexibility can negatively impact the extraction process due to the imposed tight bound. Another issue that affects many moving object pattern definitions is the constraint related to the object to monitor. Most of the previous approaches [38] consider that a pattern is defined over the same set of objects but, if we for instance consider animal migration, different objects can join (or leave) a group of objects that is moving towards a certain direction. Also in this case the data mining method needs to model some degrees of flexibility to manage such a situation.

As the data can contain noise or the user has approximate knowledge about the data itself, allowing to soften certain constraints or manage graduality in the objects that belong to a pattern can alleviate such problems.

In [35] we design and develop a fuzzy pattern mining approach to soften the time gap constraint in the field of moving object data mining. This pattern definition allows the user to define an interval time gap

(minimum/maximum) and fuzzy logic operators are introduced to retrieve patterns that match the fuzzy constraints. From an algorithm point of view, we still exploit an itemset-based representation so that GET_MOVE can be employed to gather all the fuzzy moving object patterns.

The issue related to analyze and extract object trajectories in which individuals can join (or leave) the backbone of the pattern is addressed by [37]. The proposed algorithm manages gradual moving object patterns which satisfy the graduality constraint during at least min_t timestamps. As state before, the graduality constraint allows the number of objects to increase (or decrease) while the set of remaining objects shared should be the same among the clusters.

2.2.3 Mining Representative Moving Object Patterns

Most of the researches in moving object mining are focused to extract specific trajectory patterns that differ in their characteristics with the goal to capture various aspects of the data [51, 57, 96, 37]. All these methods extract thousand of patterns resulting in a huge amount of redundant knowledge that is difficult to exploit. Spatio-Temporal databases involve complex information. Due to this complexity, studying a spatio-temporal database by employing only a single type of pattern is not enough to depict an informative picture of the data.

Motivated by these issues, we develop a Minimum Description Length (MDL)-based approach that compress spatio-temporal data leveraging different kinds of moving object patterns [36].

The proposed approach introduces a way to evaluate to which extent each extracted pattern is useful and not redundant to summarize the original spatio-temporal dataset. The MDL criterion allows to rank the set of heterogeneous patterns selecting only those that best compress the data and discarding all the useless ones.

2.3 Spatial Interaction: One more piece of the puzzle

During the last years, I moved more and more my attention towards the analysis of remote sensing data. Such source of information still belongs to the category of spatio-temporal data but, differently from the information I worked on before, it has some peculiarities: i) the spatial dimension plays a role at least as important as the temporal aspect; ii) data are mainly represented through images; iii) in most cases images are multi-bands (or hyperspectral) and they can be enriched by additional spatial information (Digital Terrain Model, etc.). Classical data mining and machine learning approaches cannot be directly applied on such data without an important pre-processing and transformation step and, such operation needs to be adapted to the particular task we would deal with.

In the field of spatio-temporal data analysis, modeling the data via graph-based representation can be beneficial to analyze information from both spatial [89] and temporal [17] point of views. From a spatial point of view, the graph structure can supply many information about how the objects

of the database are arranged while, from a temporal perspective, explicitly state the links (or relationships) among objects in a graph can help to describe and/or simulate a temporal process.

Graphs, in computer science, are an ubiquitous structure that can be easily employed to model real world information. This tool is particularly suitable to represent interactions between data [3]. Examples of graph data are social networks, gene-gene interaction networks, linked open data, document networks, knowledge graphs, etc... On the other hand, the flexibility of such a structure also allows to model data that does not naturally fit the network paradigm. The graph representation helps to highlight the interaction among the objects of the data and this interaction can be leveraged by specialized data mining and machine learning methods. For example, a textual collection can be represented as a graph where documents are nodes and a link exists between two nodes (documents) if their similarities is above a certain threshold [78]. Another similar example is supplied by the analysis of remote sensing images. After a first pre-processing segmentation step, the image segments can be represented by nodes and a link exists between two nodes if their corresponding segments spatially overlaps [32] or they have similar spectral signature [31].

2.3.1 Understanding Temporal Evolutions

During the last period I gave more and more attention, as an application domain, to the analysis of Remote Sensing data. In particular, thanks to the collaboration with experts in remote sensing analysis, we developed new data mining and machine learning approaches especially tailored for this kind of data. Recently, during the post-doctoral period of Dr. Fabio Guttler, we have started to employ graph-based analysis to model and describe temporal evolutions from time series of satellite images [32, 33]. A sketch of the process is illustrated in Figure 2.1.

Given a time series of satellite images the method performs the following steps: i) segments all the images by collecting together all the segments; ii) among the segment set, it chooses a set of reference objects maximizing the covering of the study area and minimizing the overlapping between the chosen reference objects. A reference object can come from any image of the time series; iii) For each reference object, it builds a DAG (Directed Acyclic Graph) that connects all the segments that intersect the reference object over all the images of the time series. The obtained DAG is a K-partite graph where K is equal to the number of images in the time series [32]. A direct edge exists in the DAG if the segment at level $K-1$ (segment of the image $K-1$) spatially intersects the segment at level K (segment of the image K). Figure 2.2 shows an example of an evolution graph (DAG) with the corresponding segments ordered by time stamps.

A graph can be easily exploited to describe the evolution of a zone. Given a study area (represented by a time series of satellite images), a collection of evolution graphs can be extracted. Such a set of graphs can describe the different phenomena present in the data. Most of the previous proposed methods mainly contemplate an analysis at pixel level for a reduced number of images (two or three) [42] while the novelty of our proposal lies in the use of objects instead of pixels to describe spatio-temporal evolutions.

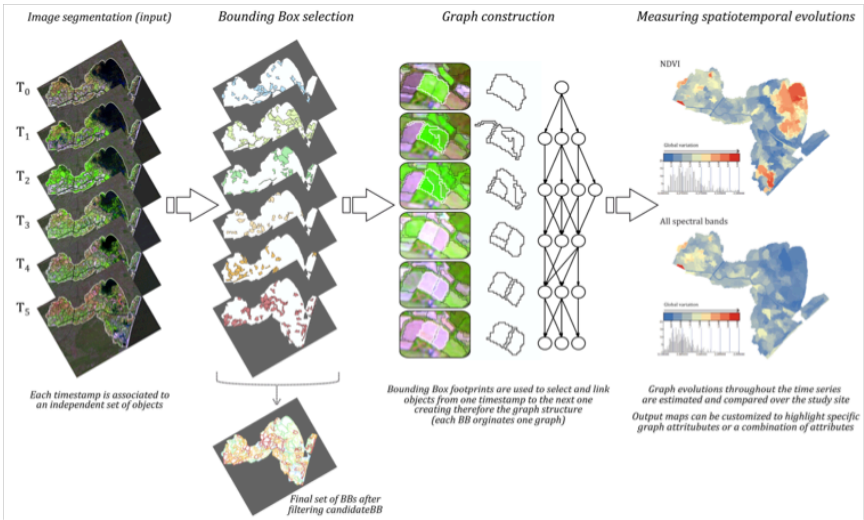


FIGURE 2.1: (Guttler et al. - to appear) The framework to extract evolution graphs from a time series of Remote Sensing Satellite Images.

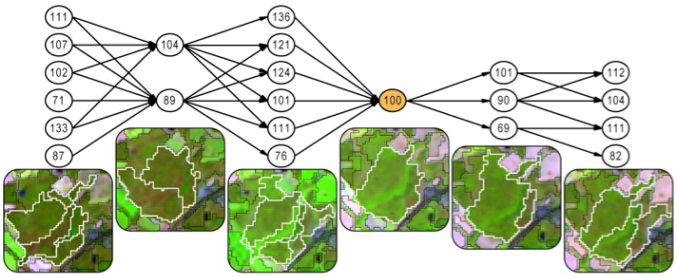


FIGURE 2.2: (Guttler et al. - to appear) An example of evolution graphs (K-partite DAG) from a time series of six images.

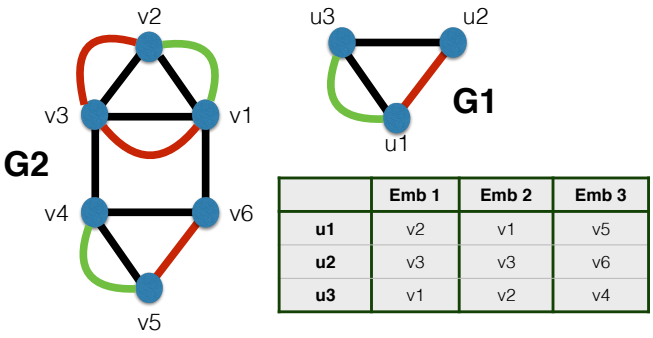
2.3.2 Manage Rich Relational Structure

Simple graph structures, sometimes, are not enough to well describe the richness of real world data. Nowadays many types of data exhibit complex relational structures where additional information in the form of multiple edges between nodes exist. Such kinds of network structures can be defined as multigraphs or multilayer graphs [71, 76, 91, 75]. They allow different types of edges in order to represent different types of relations between vertices [10, 84, 14, 52]. Many real world scenarios can be modeled as multigraphs. For instance, by considering different social networks spanning over the same set of people, but with different life aspects (e.g.

social relationships such as Facebook, Twitter, LinkedIn, etc.), we can get as many edge types as different aspects. In biology, protein-protein interaction multigraphs can be created considering the pairs of proteins that have direct interaction, physical association or they are co-localised [14]. In addition to these examples, Resource Description Framework (RDF) graphs can be naturally represented as multigraphs where the same subject/object node pair is connected by different predicates (properties) that describe different types of relationships [58].

Recently, I focused my attention to the study of such particular structures [71, 76, 91, 75] not only because multigraphs can model a wide range of application scenarios but, they can be extremely useful to model, mining and analyze spatio-temporal data. More in detail, in the context of the PhD Thesis of Mr. Vijay Ingalalli, we have designed and developed efficient methods to deal with the problem of subMultigraph iso and homomorphism [90, 91]. Both problems are known to be NP-Complete and, to some extent, the homomorphism problem is more general than the isomorphism one.

In the case of the multigraph structure the sub isomorphism test needs to take into account the topological structure but also the containment relationships between the edge set linking a pair of nodes. An example is given in Figure 2.3. In this example, we can observe two graphs: $G1$ and $G2$. The table in Figure 2.3 reports the set of embeddings of graph $G1$ in graph $G2$ with the corresponding mapping. If we consider the embedding Emb 1 the edge ($u1$,$u2$) of $G1$ matches the edge ($v2$,$v3$) because the red edge between ($u1$,$u2$) is contained among the edges (red, green) between ($v2$,$v3$). The same consideration applies for the Emb 2.



	Emb 1	Emb 2	Emb 3
u1	v2	v1	v5
u2	v3	v3	v6
u3	v1	v2	v4

FIGURE 2.3: Sub Isomorphism on Multigraph: Embeddings of $G1$ in $G2$. The table list the possible matching between nodes in $G1$ and vertex in $G2$.

Efficiently performing subgraph matching is useful for image analysis [28] as it can be used as a flexible query mechanism to answer spatial queries. The original image can be represented by a graph structure and the goal is to detect and retrieve all the portions of the images that match a certain geospatial pattern.

2.4 Remote Sensing Classification: Application to Satellite Images

Remote sensing images are an useful source of information to monitor spatio-temporal phenonmena [99]. The increasing number of projects and researches about remote sensing supplies huge amounts of data that are produced almost every day. For instance, the SENTINEL project[2] promises to capture high-resolution images every five/ten days producing a huge volume of images to process. Studying and developing suitable machine learning and data mining techniques to efficiently manage such kinds of data can be crucial for different purposes such as: food monitoring in remote regions, climate change understanding, land cover and land use classification, complex landscape description [68, 88].

2.4.1 Combine Active and Transductive Learning

Data labels are usually difficult and expensive to obtain. Standard classification techniques heavily rely on the hypothesis that a big quantity of labeled examples (training set) is available to build a predictive model. Considering the remote sensing domain the label acquisition constitutes a time and effort consuming task for the expert [23]. Classical supervised inductive classification approaches (i.e. SVM, Naive Bayes, Random Forest, etc.) require many labeled data to train the model. Also, they assume that training and test data are not available at the same time since the model they have learnt needs to be general enough to classify new unseen examples available in a near future [92]. However, in the case of remote sensing image classification, training examples are limited and all the examples (training and test) are available at the same time. Most of the time, a predictive model is learnt on a portion of the image and it is successively employed to classify the rest of the same image. **To tackle these two characteristics in land cover classification: i) data labels acquisition and ii) training and test data available at the same times, we propose in [31] a new active/transductive learning framework to cope with object-based remote sensing classification.** Transductive learning [83] belongs to the family of semi-supervised approaches. The goal of this kind of methods is to propagate information from the labeled data to the unlabeled one leveraging the availability of training and test data at the same time. These kinds of techniques offer an effective approach to supply contextual classification of unlabeled ones by using a relatively small set of labeled examples. To deal with both label scarcity and quality of training set, we couple the transductive strategy with active learning with the goal to improve accuracy performance and supply a valid alternative to standard classification techniques usually employed in remote sensing domain such as Support Vector Machines and Random Forest [41].

2.4.2 Time Series

Although data from satellite images are very useful for monitoring land surface, the large quantity of spatio-spectro-temporal measurements stored

[2] http://www.satsentinel.org/

by the instruments limits the usefulness as sources of information. In recent years, research on spatio-temporal databases has consequently increased alongside research on mining such data [11]. **In [74] we deal with the classification of remote sensing time series data with the purpose to characterize land use in the northern fringe of sub-Saharan Africa**. In this work we put a major stress on the temporal dimension. More in detail, we developed a data mining methodology to extract multidimensional sequential patterns to characterize temporal behaviors. In the same spirit as [66], we used the extracted multidimensional sequences to build an associative classifier, and show how the patterns help to discriminate among the classes. We evaluated our technique using a real-world dataset with the purpose to automatically recognize the land use of a certain area.

2.5 Data Stream Classification: Deciding when update the model

Many real world applications continuously generate huge amounts of data, such as web logs, sensor networks, business transactions, etc. These *data streams* [2], due to the big volumes of information they contain, pose serious issues for the research community in order to extract useful and up-to-date knowledge in real-time. Due to its intrinsic temporal dimension, the information available in data streams can change and evolve over time. More precisely, this phenomenon impacts on the performance of any supervised (or unsupervised) model learnt over these evolving data: previous models may not be suitable for newly incoming data [2]. Therefore we need to adapt models both quickly and accurately.

2.5.1 Active Learning

Learning predictive models on streaming data implies having continuous access to the true values of the target variable (the true class labels) of every incoming example. This labeling phase is usually an expensive and tedious task for human experts. Consider, for example, textual news arriving as a data stream. The goal is to predict if a news item will be interesting for a given user at a given time. The interests and preferences of the user may change over time. To obtain training data, news items need to be labeled as interesting or not. This requires human labor, which is time consuming and costly. For instance, Amazon Mechanical Turk[3] offers a marketplace for intelligent human labeling. Labeling can also be costly or practically unfeasible because it may require expensive, intrusive or destructive laboratory tests. The labeling problem in standard machine learning scenario is well known [87] and it is mainly addressed through techniques that guide the construction of the training set by the needs of the predictive models. Such kinds of techniques belong to the family of active learning [26]. **To address this important issue in the context of data streams, during the period spent at the University of Waikato, in collaboration with Prof. Bernhard Pfahringer and other colleagues, we developed new active learning strategies especially tailored to efficiently learning predictive models on evolving data streams [49, 47].** All the proposals we developed are based

[3] https://www.mturk.com

on the idea that important instances to sample lie in a high density partition of the data space. In [49] we instantiate this idea by exploiting clustering approaches in order to estimate the density around each point while in [47] we estimate the density in an online manner, a sliding window mechanism allows to quantify the importance of a point considering the density of its nearest neighbors.

2.5.2 Categorical Change Detection

Due to its intrinsic temporal dimension, the information available in data streams changes and evolves over time. In particular, different types of changes may happen in the stream. For instance, classes or concepts that can be underrepresented during a short period can become overrepresented after a longer period. Most of the time, a common assumption made by many research works is to consider only the aposteriori probability of the class given the data [27, 9]. This formulation of the change detection task does not exploit information coming from the underlying data distribution. Another issue is that, in real world applications, data is heterogeneous and often can be represented over set of categorical attributes as well as numerical ones. In the last decade, lots of approaches have been defined to monitor classification accuracy as evidence or an indication for change in streams of numerical data [9] but, few approaches dealing with the same problem for categorical evolving data streams [19]. **In [48], we tackle this issue and we propose a new change detection approach devoted to retrieve changes in categorical evolving data streams.** The proposed approach detects and highlights changes in categorical data streams in a fully unsupervised setting. It works in the batch scenario: when a new batch arrives, firstly the algorithm summarizes the block through some statistics and successively performs a statistical test to evaluate if a change happens in the data distribution, or not. The developed algorithm supplies a segmentation approach that can also work with other statistics. This means that it can be coupled with any other measure we want to monitor.

2.6 Other Research Activities in Data Science

During the last years, I devoted a major part of my researches to analyze spatio-temporal data but, in order to enrich my methodological background in the field of data mining and machine learning, we studied, conceived and developed, in partnerships with other colleagues, approaches to manage different kind of data such as textual, categorical and linked open data. The motivation related to these researches is that, a methodology we can apply on textual data (i.e. supervised classification, clustering algorithm, etc.) can be adapted and reused, to some extent, for spatio-temporal data analysis. As example, in the field of multilingual document classification we experimented and apply transductive based methods [78] and, successively, we extend and adapt the same strategy to perform active transductive classification in the context of object oriented Remote Sensing analysis [31]. Study and design general data mining and machine learning methods allows me acquire a good knowledge on how customize each of such strategy w.r.t. the particular domain to investigate.

2.6.1 Clustering and Co-Clustering Unstructured data

Most of the techniques developed during my PhD training were devoted to unsupervised analysis and in particular to cluster data. During the last years, we continue, with the colleagues I am working with, to conceive and propose new clustering techniques . Most of the proposed techniques were devoted to analyse textual information.

The choice of this source of information as type of data is due to its ubiquitous nature and to the fact that it is generally not so difficult to retrieve annotated document collections. The availability of labeled data facilitates the evaluation of the different methods.

As a direct extension of my PhD work, we have developed new co-clustering techniques for dynamic textual sources of information [73] and co-clustering techniques for heterogeneous data [50]. In the domain of dynamic textual data we developed approaches able to incrementally cluster streams of text supplying a hierarchical organisation of such documents. The hierarchical organisation allows to easily explore and browse the content of the evolving document collections. In the same period, we also focused our attention to develop clustering methods that allow the grouping of entities that have an heterogeneous representation. For instance, in [50] we show how entities that can be described, at the same time, by both visual (images) and textual (captions) information can be jointly analysed in an unsupervised way.

Data heterogeneity is an important characteristic of modern sources of information. The same information can be described by different representations or by different media. An example of such heterogeneity, in the field of text analysis, is supplied by multilingual document collections [79] in which the same information can be available in different languages. **During the co-supervision (with Dr. Andrea Tagarelli) of the PhD Thesis of Dr. Salvatore Romeo, we designed and developed on unsupervised and semi-supervised machine learning framework to analyse multilingual document collections [79, 78].** To tackle the issue of language heterogeneity, we leverage knowledge-based resources [69] to obtain a common representation for collection of multilingual documents. Once the new representation was obtained, we have developed new data mining techniques to automatically categorise multilingual textual information [79]. In the context of multilingual textual clustering we modeled the documents by a tensor representation and we successively factorized such tensor to find a low dimensional embedding of the original data that helps the clustering process. In the context of semi-supervised textual categorization [78] we have employed a graph-based transductive learner to propagate label information from labeled documents (written in a certain language) to unlabaled ones (written in the same or another language). The propagation process supplies the final classification result.

2.6.2 Semi-Supervised Learning in Categorical Data

Supervised and Unsupervised settings, from a machine learning perspective, are two extreme cases in which, from one side, we have labels for all the training data and, from the other side, we do not have class information for any of the objects we would analyse. In many real world tasks the

situation is less strict and, we can have situation in which only a portion of the data has associated labels. We can image a scenario in which we need to classify home pages vs. non home pages. In this context we have a good knowledge about what is an home page but, strictly define what is not an home page can be difficult. Examples of such scenario are semi-supervised anomaly detection [20] and positive and unlabeled learning [56]. Due to the variety of contexts in which semi-supervised analysis can appear, this is also valid in the context of remote sensing image analysis [31]. **More in detail, in [46] we recently proposed a semi-supervised anomaly detection for categorical data where a model is learnt on "normal" data and then the learnt model is employed to rank anomalies entities. For the problem of positive and unlabeled learning, we propose in [45] an approach that firstly computes a model for the positive class, secondly a set of representative examples for the negative class is detected and finally a discriminative model is built considering both positive and negative instances.**

The peculiarity of the strategies we proposed is related to the data they manage: all the examples are represented by only categorical variables (any numerical variable can be discretized to obtain a categorical one). The problem to employ distance-based machine learning methods on categorical dataset is related to the notion of distance measure [20]. Unlike numerical attributes, it is difficult to define a distance between pairs of values of a categorical attribute, since the values are not ordered. This underlines the fact that adapt distance-based machine learning methods to manage categorical information is challenging.

2.7 Conclusion

This chapter resumes the major researches I developed in the last years in collaborations with my colleagues. Most of the themes we addressed are related to the design of general data mining and machine learning techniques that can be applied on spatio-temporal or complex data. Each of the addressed topic can supply ideas or hints for possible follow-ups. For instance, regarding the study of moving object data, a deeper investigation can be made on how moving object patterns can be summarized and how such results can be presented. In the data stream scenario, a possible point to address could be the study of how coupling active learning and semi-supervised learning or how unsupervised change detection techniques can help unsupervised learning (such as clustering) to extract useful information from evolving data. Any of the previous macro topics can supply ideas and point out possible research tracks in the domain of spatio-temporal, textual or categorical data analysis.

In the next chapter I will try to sketch some of the research directions I would investigate that are deeply connected to my current research in spatio-temporal data analysis.

Chapter 3

Perspectives

This chapter draws perspectives related to my on-going research on spatio-temporal data mining. I would discuss in which direction my research activities can evolve in the next four, five years. All the perspectives presented in this chapter are related to students I am co-supervising and projects I am involved in.

3.1 Research Objectives

In the last years, from the time I joined the TETIS laboratory, I concentrated my research efforts toward the analysis of dynamic data in which the temporal dimension plays a major role. As I discussed in the previous chapter, I studied, designed and developed, with researchers I worked with, new approaches in both pattern mining and machine learning fields to manage, in particular, the temporal dimension present in spatio-temporal databases. The complexity of such data is given by the temporal but also by the spatial information it contains. This is why, my next research objectives will focus on increasing my knowledge through the study and the development of methods that explicitly deal with spatial autocorrelation inside the data.

My research background ranges from supervised and semi-supervised machine learning methods to cluster analysis, to the design of pattern mining approaches to mine and extract knowledge from databases. In the current data science literature, many methods to manage spatial information already exist [11]. My main objective is to design, develop and implement new approaches development of new approaches to deal with the particular kind of data I am focusing on: Time Series of Remote Sensing Data. To pursue this goal, in the future, the next PhD fellowships I will supervised, they will be focused on the analysis of general spatio-temporal data mining techniques considering the Remote Sensing field as privilegiate domain of application.

The PhD students I am currently co-supervising, are mainly focusing on methods to mine and learn from Remote Sensing (RS) Data. Particularly, the different on-going thesis in which I am involved, they also reflect my different research backgrounds (pattern mining, cluster analysis and supervised learning).

3.2 Graph-based approaches to analyze spatial information

General graph data mining approaches [101, 53, 24] traverse the search space of graph patterns and, once a pattern is generated, test its support. The algorithms can work in a transactional setting [101] (the database is constituted by a collection of graphs) or extract frequent graphs in a single graph setting [53, 24] (the database is constituted by only one big graph). The most time consuming and crucial operation in such approaches is the subgraph isomorphism test [15] that allows to retrieve all the embeddings (occurrences) of a graph pattern in the graph database. Multigraph pattern mining seems similar to general graph mining but it has its peculiarity. The main difference relies in the subgraph isomorphism test [90]. During the on-going PhD Thesis of M. Vijay Ingalalli, co-supervised in collaboration with Prof. Pascal Poncelet (LIRMM), we have developed a method to perform subgraph isomorphism test for multigraph [90]. The next step of this research involves the design and the implementation of a frequent multigraph mining approach. The approach will leverage techniques and tricks already proposed in the general domain of graph mining but it will introduce specific pruning strategies tailored for the multigraph structure.

In the remote sensing context we can take, as a toy example, the extraction of complex landscape interactions. Given a satellite image we can easily model the image content as a multigraph. After image segmentation, for each segment we can compute a set of characteristics induced by its radiometric attributes. This results in a vector of numerical features. The set of segments is the set of the nodes of the multigraph and two nodes are linked each other if they are spatially adjacent. More in detail, we can have a different edge type for each of the segment's feature. In this way, an edge of type t_i exists between two objects if they are spatially adjacent and they have similar values w.r.t. feature f_i. An example of this process is reported in Figure 3.1 where, for the sake of clarity, a geographical area is segmented, numerical features are computed and the mulitgraph is built following the previous strategy. Different edge types are represented by different colors.

Once the image is represented by a multigraph we can extract frequent sub(multi)graphs in order to study how segments spatially interact with each other. Frequent patterns can be used to highlight recurrent landscape interactions. Explicitly considering different type of edges can supply more fine information to experts about the physical phenomenon and what the study area contains. Frequent patterns can be employed directly or they can be used, for instance, to feed statistical simulation process. In environmental analysis it is common to design mathematical models to simulate physical and natural evolutions. Build such a model is a time consuming task if we do not know what happens in the area we would study. Frequent sub(multi)graphs patterns can be used to guide the construction of such models.

3.2.1 Select Interesting Multigraph Patterns

Pattern mining methods that only extract frequent patterns can potentially produce a huge number of patterns that can be hardly analyzed by human experts, hence limiting the usefulness of such tools. To cope with this issue,

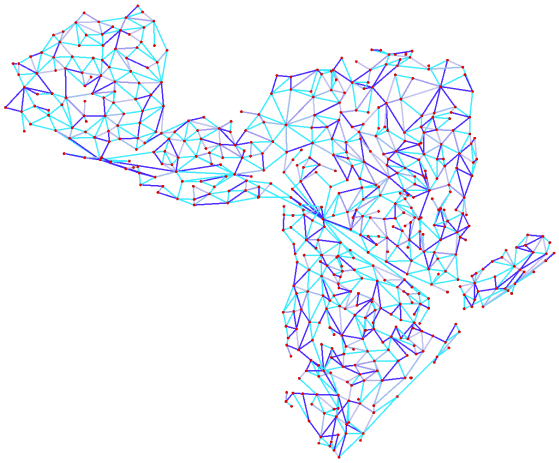


FIGURE 3.1: Example of multigraph representation for a segmented remote sensing image. The different colors indicate different type of edges.

three different families of approaches can be found in literature to extract interesting patterns, during the mining algorithm or as post-processing step. This categorization only considers the big trends in the pattern mining field and it does not want to be exhaustive.

The first family of approaches relies on compression-based measures such as MDL [94] or entropy [85]. Such methods filter out only a subset of interesting patterns selecting those ones that better compress the original database or that are unexpected considering their constituents.

The second family of approaches leverages statistical models in order to represent the underlying distribution of the patterns. In this case a null-model is built and patterns that do not fit this null-model are selected as interesting due to their exceptionality [86, 60].

The third family of approaches always exploit statistical theory but they extract frequent patterns through a stochastic process. More in detail a sampling algorithm (usually based on Markov Chain Monte Carlo Methods) is designed and patterns are sampled from the distribution. Sampled patterns are characterized as interesting [77, 80, 95, 39].

In the graph mining field, most of the strategies to extract interesting patterns relie on the third family of approaches.

The first family of methods mainly selects interesting patterns as post-processing. First of all the frequent patterns are generated and then the relevant ones are selected as post-processing. In the case of graph patterns, conceiving and designing compression measures to post-process the set of mined patterns is challenging and more difficult w.r.t. equivalent measures for itemsets or sequences. On the other hand, compressing the original database (or considering if a superpattern is more interesting than a sub-pattern) involves an heavy use of the subgraph isomorphism procedure that can drastically increase the time of such methods.

The second family of methods to filter out interesting patterns needs the construction of a null-model of the data distribution and the null-model is exploited in order to prune the search space. Till now, from the best of my

knowledge, no approaches for graph mining towards this direction exist and, for sure, it will be interesting to understand the feasibility of this strategy in the context of multigraph pattern mining. The challenge related to this point are i) the construction of a null-model to represent the topological structure of the multigraph ii) the design of an efficient algorithm to navigate the search space according to the proposed null-model.

The third family, until now, seems to be more pertinent in the graph mining domain [39]. Probably, it is due to the fact that sampling patterns instead of exhaustively traverse the whole search space drastically decrease the running time and allows graph mining approaches to scale up on bigger datasets producing a restricted amount of patterns that can be easily investigated by a domain expert.

Concerning the extraction of interesting multigraph patterns, the research can address the study of already proposed strategies coming from the last two families of approaches and how to generalize such groups of methods to address the selection of interesting multigraph patterns.

Orthogonally to the selection of interesting patterns, another research direction can be devoted to introducing and designing constrained based pattern mining algorithm especially tailored for multigraph data. More the structure becomes complex (itemset, sequence, graph, multigraph) more constraints we can define in order to filter out patterns that meet such requirements [104]. Designing efficient algorithm to manage constraints in multigraph is challenging and it can constitute another fruitful research direction.

3.2.2 Multigraph Rules

Another possible research direction can be represented by the study of techniques to extract multigraph rules. Similar to association rule mining [21], extract rules on multigraph databases can help to better characterize the information contained in the database and to understand cause-effect relationships leveraging frequent patterns. Preliminary works were done in the context of graph mining for evolving graphs [17] and recently, some works suggest to relax the constraints about the topology and, successively, mining node labels that frequently occur near each other [40]. Such rules could also be useful from a database point of view to extract conditional dependencies to summarize and describe multigraph databases.

3.2.3 Approximate Multigraph Mining

Most of the previous follow-ups assume that the sub(multi)graph isomorphism task performs exact matching. Many times, in real world scenario approximation is necessary in order to deal with possible noise or uncertainty present in the data [67, 6, 82]. Approximation can be performed at information level [6, 18] (label of the nodes) or a topological level [103](modify the sub isomorphism algorithm to find approximate embeddings). In the multigraph context, one more piece of information to manage will be the presence of multiple edge types that can be modeled as an edge with an associated itemset as label. Depending on which level of approximation we want to deal with, we need to redefine primitive operations or only data mining algorithms. Logically, redefining primitive operations will take

more time than modifying the graph mining algorithms but, it will allow to acquire more knowledge about the whole graph mining process.

3.2.4 Applications on other domains

The design and the development of multigraph pattern mining approaches is primarily motivated by the abundance of spatio-temporal data that can be modeled as graphs and also by multigraphs. The toy example I supplied in Section 3.2 about image analysis through multigraph pattern mining is only one example but another example can be the extraction of colocation patterns [102] (commonly employed in spatial data mining) from a multigraph that represents how objects are spatially arranged with a more fine-grain description about how they are interconnected. Due to the wide range of databases that can be represented as multigraphs, the proposed perspectives are not only limited to the analysis of spatio-temporal data but they can be beneficial to mine information coming from other domains. Among all the possible domains the previous proposals could affect, the analysis of knowledge graphs seems to be one popular example [70, 100, 59]. Knowledge graphs are structures that supply a network representation of knowledge where the nodes are objects and links between objects represent some kind of relationships. In knowledge graphs, the same pair of nodes can be linked by different edge types resulting in a multigraph structure. Examples of such knowledge graphs are Yago [64], DBPedia [55] and FreeBase [12]. Data Management and Mining techniques able to efficiently deal with multigraph data can be useful to extract information that could be reused by high-level approaches to reason [25] about the underlining knowledge. The same approaches can be employed over other domains in which the multigraph structure can appear such as bioinformatics and social network analysis [10, 14].

3.3 Summarize Temporal Evolutions in RS Time Series

Among the different researches I would pursue, the analysis of time series of remote sensing data still plays an important role supplying interesting scenarios to develop new data mining and machine learning techniques. Due to the new research programs that promise, in the next years, to drastically increase the volume of data acquired by satellite sensors, the field of remote sensing time series analysis will probably get more and more attention from the research community. Practically, this huge volume of data will pose new challenges in order to be analyzed efficiently [62, 63]. This research track is currently related to the thesis subject of Mme Lynda Khiali, a PhD student I am co-supervising with Dr. Maguelonne Teisseire (IRSTEA).

3.3.1 Summarize evolutions in RS Time Series

During these researches we would design and implement new methods to study how the entity evolves in remote sensing time series leveraging the graph representation we previously introduced. The work we previously did was related to the analysis of time series of images spanning over

only one year and it was limited to extract as many evolution graphs as the selected reference objects. One extension we would work towards is the summarization of such set of graphs. Among the possible extracted graphs, many of them can represent the same (or similar) information producing some kinds of redundancy. To address this issue a solution can be the use of clustering techniques [43] with the purpose to summarize the collection of graphs. In order to cluster (and summarize) such kind of information we need to understand how to evaluate the distance between such graphs [81]. This is a fundamental operation in any distance-based data mining approach (i.e. clustering). How to define a suitable distance between graphs is still challenging [81] as it depends from i) the characteristics of the graph structure we analyze (labeled vs unlabeled, DAGs vs general graphs, weighted vs unweighted, etc..) and ii) the task we would accomplish (summarization, diversity, indexing, etc...).

Once the distance measure and clustering algorithm are available, some efforts can be made in order to introduce more supervision in the summarization process via end-user interactions. Due to my previous experiences, a way to introduce a limited amount of user feedback in the mining process could be the combination of clustering [43, 73, 79] and active learning [26]. Until now, active learning was mainly exploited in supervised scenarios (classification tasks) but few works start to appear in the literature about how to combine clustering approaches and active learning [26]. Due to the peculiarity of remote sensing data in which spatial and temporal information plays a crucial role, clustering such kinds of data requires appropriate methods [16] and the active clustering approaches recently proposed [72, 1, 98] completely ignore the spatial and temporal dimensions in their sampling process. The research track related to active clustering methods for spatio-temporal data is still challenging and it can constitutes, from my point of view, a valuable field of research to investigate in order to supply user-oriented data summarization.

3.3.2 Mining Episodes in Evolution Graphs

During the work we did in [32, 33] we worked on time series of remote sensing data that spans over one year in order to describe the phenology of the studied area. In the context of climate changes we are interested in following natural phenomena over ten, twenty years or more. To this purpose, we need to analyze multi-annual time series of remote sensing images. Considering the seasonality of natural phenomena, if we only consider yearly time series, most of the phenomenon appear only once and the technique we have proposed is easily applicable to such scenarios. When the time series involves long periods our approach can have some issues due to the preliminary assumptions it made [32]. In the case of multi-annual time series what we can expect is to find recurrent signals that can reproduce themselves with some approximation. In the pattern mining field, the study of recurrent events that can appear in a (possibly infinite) sequence of data goes under the name of episode mining [97]. The goal of such techniques is to extract patterns of evolution that can be recurrent along the whole sequence of data. While in the context of general episode mining, the data arrives sequentially and the stream is produced somewhere else, in the context of multi-annual remote sensing images we can imagine a scenario

in which, applying the same process we already did in [32], we can build evolution graphs that cover multi-annual time series. Once the evolution graphs are built, successively (or during the construction of graphs), we can identify recurrent substructures that can allow, for instance, to segment the whole sequence of data or summarize again the whole dataset. Applying the graph extraction step on a time series of satellite images results in a set of evolution graphs. Such collections of graphs can be seen as a database to be mined itself. Defining data mining algorithms in order to extract frequent and/or interesting episodes from a database of graph sequences is challenging and needs to consider the particular information that such structures represent (recurrent events, spatial covering, anomaly behavior, etc...).

3.4 Exploit Spatial Autocorrelation in remote sensing analysis

Another point I am interesting in is related to supervised and semi-supervised learning approaches devoted to classification purposes [66, 45, 46]. In the last period, I started to develop such methods in the context of satellite image classification [74, 31]. Automatic classification methods (supervised and semi-supervised ones) are important in the context of remote sensing analysis where an image can contain thousands of pixels and/or thousands of objects to process. Examples are satellite image classification that considers land use or land cover soil classes [74]. Recently, in the field of machine learning, Deep Learning methods stand up from the crowd of classification methods underlining that such strategies are able to heavily outperform state-of-the-art approaches [54]. Among this family of methods, Deep Convolutional Neural Networks (CNN) show impressive performance in the field of image classification [54] thank to the Convolutional layer that allows to capture, to some extent, spatial autocorrelation among the pixels of the image. A classical CNN architecture is reported in Figure 3.2. We can observe different types of layers, the first one represents the input data, then we can note the portion of the network dedicated to the convolutional operation and, finally, the last layers represent a fully connected Multi Layer Perceptron that produces the final classification.

3.4.1 Deep Learning in Heterogeneous and Incomplete Data

In 2015 I started to investigate Deep Learning methods and, still in the same year, I started the co-supervision with Dr. Marc Chaumont (LIRMM) and Dr. Gerard Subsol (LIRMM) of the PhD Thesis of M. Lionel Pibre on the "Detection of urban objects from heterogenous remote sensing data sources". This work will be devoted to the use of deep learning techniques on heterogeneous remote sensing data to detect objects (i.e. trees) in urban areas. As an objective, we would overpass some limitations of standard supervised approaches on heterogenous multi source data.

Standard classification algorithms need a training phase before being employed on new unseen examples (test data). A challenge in this field can be the development of new deep learning architectures to manage heterogeneous data coming from different sensors (optical sensor, satellite images,

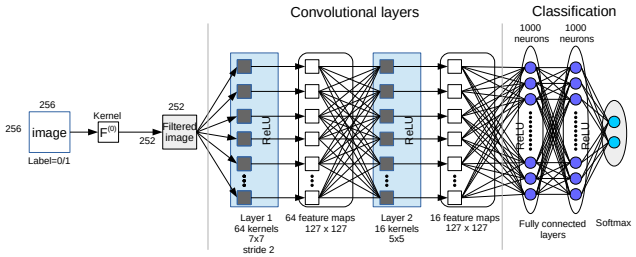


FIGURE 3.2: An example of Convolutional Neural Network (CNN) architecture with a first fixed Kernel to preprocessing the input image, two convolutional layers to manage image autocorrelation, two fully connected layers to prepare the prediction step and, finally, a SoftMax layer to perform the final classification.

drone images, Radar images, etc...) in a setting where training and test data cannot fit the same format and, commonly, the test set is poorer (in terms of information) than the training set. In this case we can exploit previous approaches proposed in the field of deep and representational learning [54, 8] in order to obtain some intermediate representation coping with partial information. Such intermediate representation can be obtained employing some kind of autoencoder tool [93]. Successively, the deep learning system can be trained on such intermediate representation and, once a new unseen example is available, first of all it will be transformed employing an autoencoder and then the transformed instance can be the input of the predictive model to get the final classification.

During the preliminary experiments we conducted, we observed that the results are heavily influenced by the type of activation function employed in a particular layer of the CNN. Studying some kind of adaptive way to choose which activation function could be used in which layer can be useful in general. For instance, instead to use a single activation functions [8] (ReLU, eLU, etc...) for each layer, maybe a weighted linear (or not linear) combination of such functions can allow the network to automatically learn which one is more suitable at which point of the architecture. All these perspectives can be developed considering as an application scenario the analysis of remote sensing images where, only in the last two years, deep learning approaches started getting increasing attention [61, 65].

3.4.2 Deep Learning for image data archives

As well underlined by [54], one of the possible future directions of deep learning is the use of such models in an unsupervised way. Recently, due to the rapid evolution of satellite system, large-scale remote sensing image data archives are more and more available and they need practical tools to organize and retrieve such information. An exhaustive search through linear scan over such archives is really time-consuming and not practically reasonable in real world applications. Recently, [22] proposes an hashing-based scalable remote sensing image search systems to overcome this problem employing kernel methods to hash such images. An interesting point

to investigate could be the use of Deep Learning architecture in order to compress remote sensing satellite images performing some kind of Local Sensitivity Hashing [29] to speed up the search and retrieval process. Such hashing can be successively indexed by some tree structure and the index can support approximate similarity queries. Leveraging the ability of CNN to model spatial auto-correlation, the obtained compressed representation can implicitly incorporate spatial information. The encoding supplied by the Deep learning strategies can be learnt in both unsupervised or (partially) supervised ways in order to supply a more efficient image search engine. Combining Deep Models (with a representational purpose) with information retrieval and database techniques to manage and query archive of remote sensing satellite images can be an interesting field of research. In the past, examples in which machine learning techniques are employed to optimize and ameliorate database and information retrieval systems have already pointed out their usefulness [5, 4].

3.4.3 Mixing Convolutional and Recurrent Neural Networks

Among the perspectives in the Deep Learning fields, listed by [54], the improvement of current deep methods to deal with dynamic systems or sequential inputs will be one point to address in the near future. The particular neural architecture devoted to manage sequential and temporal information is called Recurrent Neural Network (RNN). RNNs are valuable tools that started to demonstrate their interest to classify and compress sequential data [30]. RNNs manage an input sequence one element at a time, maintaining in their hidden units a 'state vector'. Such a vector implicitly represents the information about the history of all the past elements of the sequence. RNNs, once unfolded considering the time dimensions, can be seen as a very deep feedforward networks in which all the layers share the same weights. One of the major problem of such architecture was the training phase over many time steps, in which, the network can typically explode. Recent advances in the domain of RNNs proposed techniques such as Long Short-Term Memory (LSTM) and its variants [30]. In order to deal with the problem to remember too long events, such approaches have hidden states employed as memory. The role of these hidden states is to propagate the same information from a time stamp to the next one. The improvement of such architectures is related to the ability to learn when such memories can be re-initialized or not.

In the context of time series of satellite images, as previously discussed, both spatial and temporal aspects play an important role. Such characteristics are crucial to understand the underlying behavior for both classification and summarization purposes. CNNs show impressive performance for image analysis while RNNs demonstrate their ability to model temporal data. A possible research direction can be the combination of these two models in the context of remote sensing data. What can be proposed is an hybrid architecture able to learn, at the same time, convolutional filters that are related to particular portions of the input sequences exploiting the ability of RNNs to model recurrent phenomena.

3.5 Conclusion

The first law of geography tells us that "everything is related to everything else but nearby things are more related than distant things". Such a characteristic is also known as the spatial autocorrelation. Therefore, the widely used i.i.d. assumption in data mining is too strong when analyzing spatial data. New methods and modeling techniques are needed to tackle the spatial heterogeneity and the spatial relationships (such as topological relationships, directional relationships, etc.), which are unique to spatial data. Spatio-temporal data are further temporally dynamic, which requires explicit or implicit modeling of the spatio-temporal autocorrelation and constraints to achieve good prediction performance[1]. This is why, in the past, I concentrated my effort on spatio-temporal data and this is also why I would continue, in a near future, to investigate new data mining and machine learning approaches for Spatio-Temporal data with more emphasis on the spatial component.

Le me conclude by stating the following observation: research is an active process that involves as well junior as senior researchers and as well PhD as PostDoc. From my reduced experience, research is a collective activity in which people collaborate with each others sharing experiences and new points of view about the same task. My past research was influenced by people I worked with during visiting periods, conferences, and project collaborations. Likewise, my future research will be influenced by new people I will meet in this never ending trip that is research.

[1] http://researcher.watson.ibm.com/researcher/view_group.php?id=4152

Appendix A

Mining Representative Moving Object Patterns

Mining Representative Movement Patterns through Compression

Phan Nhat Hai, Dino Ienco, Pascal Poncelet, and Maguelonne Teisseire

[1] IRSTEA Montpellier, UMR TETIS - 34093 Montpellier, France
{nhat-hai.phan,dino.ienco,maguelonne.teisseire}@teledetection.fr
[2] LIRMM CNRS Montpellier - 34090 Montpellier, France pascal.poncelet@lirmm.fr

Abstract. Mining trajectories (or moving object patterns) from spatio-temporal data is an active research field. Most of the researches are devoted to extract trajectories that differ in their structure and characteristic in order to capture different object behaviors. The first issue is constituted from the fact that all these methods extract thousand of patterns resulting in a huge amount of redundant knowledge that poses limit in their usefulness. The second issue is supplied from the nature of spatio-temporal database from which different types of patterns could be extracted. This means that using only a single type of patterns is not sufficient to supply an insightful picture of the whole database.
Motivating by these issues, we develop a Minimum Description Length (MDL)-based approach that is able to compress spatio-temporal data combining different kinds of moving object patterns. The proposed method results in a rank of the patterns involved in the summarization of the dataset. In order to validate the quality of our approach, we conduct an empirical study on real data to compare the proposed algorithms in terms of effectiveness, running time and compressibility.

Keywords: MDL, moving objects, spatio-temporal data, top-k, compressibility.

1 Introduction

Nowadays, the use of many electronic devices in real world applications has led to an increasingly large amount of data containing moving object information. One of the objectives of spatio-temporal data mining [5] [10] [6] is to analyze such datasets for interesting moving object clusters. A moving object cluster can be defined as a group of moving objects that are physically closed to each other for at least some number of timestamps. In this context, many recent studies have been defined such as flocks [5], convoy queries [7], closed swarms [10], group patterns [15], gradual trajectory patterns [6], traveling companions [13], gathering patterns [16], etc...

Nevertheless, after the extraction, the end user can be overwhelmed by a huge number of movement patterns although only a few of them are useful. However, relatively few researchers have addressed the problem of reducing movement pattern redundancy. In another context, i.e. frequent itemsets, the Krimp algorithm [14], using the minimum description length (MDL) principle [4], proposes to reduce the amount of itemsets by using an efficient encoding and then provide the end-user only with a set of informative patterns.

In this paper, we adapt the MDL principle for mining representative movement patterns. However, one of the key challenges in designing an MDL-based algorithm for

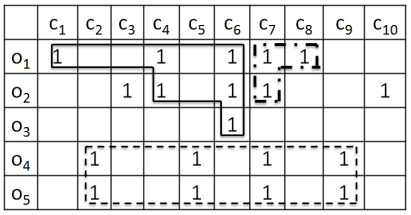


Fig. 1. An example of moving object database. Shapes are movement patterns, o_i, c_i respectively are objects and clusters.

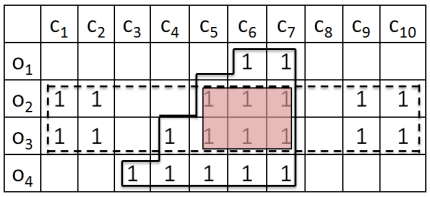


Fig. 2. An example of pattern overlapping, between closed swarm (dashed line rectangle) and $rGpattern^{\geq}$ (step shape), overlapping clusters are c_5, c_6 and c_7.

moving object data is that the encoding scheme needs to deal with different pattern structures which can cover different parts of the data. If we only consider different kinds of patterns individually then it is difficult to obtain an optimal set of compression patterns.

For instance, see Figure 1, we can notice that there are three different patterns, with different structures, that cover different parts of the moving object data. If we only keep patterns having a rectangular shape then we lose the other two patterns and viceversa.

Furthermore, although patterns express different kinds of knowledge, they can overlap each other as well. Thus, enforcing non-overlapping patterns may result in losing interesting patterns. For instance, see Figure 2, there are two overlapping patterns. Krimp algorithm does not allow overlapping patterns then it has to select one and obviously loses the other one. However, they express very different knowledge and thus, by removing some of them, we cannot fully understand the object movement behavior. Therefore, the proposed encoding scheme must to appropriately deal with the pattern overlapping issue.

Motivated by these challenges, we propose an overlapping allowed multi-pattern structure encoding scheme which is able to compress the data with different kinds of patterns. Additionally, the encoding scheme also allows overlapping between different kinds of patterns. To extract compression patterns, a naive greedy approach, named NAIVECOMPO, is proposed. To speed up the process, we also propose the SMARTCOMPO algorithm which takes into account several useful properties to avoid useless computation. Experimental results on real-life datasets demonstrate the effectiveness and efficiency of the proposed approaches by comparing different sets of patterns.

2 Preliminaries and Problem Statement

2.1 Object Movement Patterns

Object movement patterns are designed to group similar trajectories or objects which tend to move together during a time interval. In the following, we briefly present the definitions of different kinds of movement patterns.

Database of clusters. Let us consider a set objects occurring at different timestamps. A database of clusters, $C_{DB} = \{C_{t_1}, C_{t_2}, \ldots, C_{t_m}\}$, is a collection of snapshots of the moving object clusters at timestamps $\{t_1, t_2, \ldots, t_m\}$. Given a cluster $c \in C_{t'} (\in C_{DB})$, $t(c)$ and $o(c)$ are respectively used to denote the timestamp that c is involved in and the set of objects included in c. For brevity sake, we take clustering as a preprocessing step.

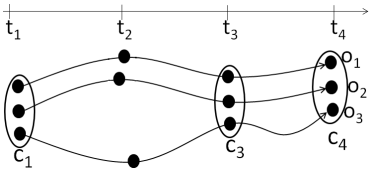


Fig. 3. An example of closed swarm.

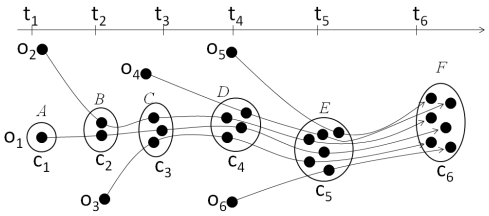


Fig. 4. An example of rGpattern.

After generating C_{DB}, the moving object database (O_{DB}, T_{DB}) is defined such as each object $o \in O_{DB}$ contains a list of clusters (i.e. $o = c_1 c_2 \dots c_m$) and T_{DB} stands for the associated timestamp. For instance, Figure 1 presents the database O_{DB} and object o_1 can be represented as $o_1 = c_1 c_4 c_6 c_7 c_8$.

From this set different patterns can be extracted. In an informal way, a closed swarm is a list of clusters $cs = c_1 \dots c_n$ such that they share at least ε common objects, cs contains at least min_t clusters and cs cannot be enlarged in terms of objects and clusters. Note that there are no pairs of clusters which are in the same timestamps involved in cs. Then a closed swarm can be formally defined as follows:

Definition 1 *ClosedSwarm[10]. A list of clusters $cs = c_1 \dots c_n$ is a closed swarm if:*

$$\begin{cases} (1) : |O(cs)| = |\bigcap_{i=1}^{n} c_i| \geq \varepsilon. \\ (2) : |cs| \geq min_t. \\ (3) : \nexists i, j \in \{1, \dots, n\}, i \neq j, t(c_i) = t(c_j). \\ (4) : \nexists cs' : cs \subset cs', cs' \textit{ satisfies the conditions (1), (2) and (3).} \end{cases} \tag{1}$$

For instance, see Figure 3, $cs = c_1 c_3 c_4$ is a closed swarm with $min_t = 2, \varepsilon = 2$. Similarly, in Figure 1, we also have $cs = c_2 c_5 c_7 c_9$ is a closed swarm. A convoy is a group of objects such that these objects are closed each other during at least min_t *consecutive* time points. Another pattern is group pattern which essentially is a set of disjointed convoys which are generated by the same group of objects in different time intervals. In this paper, we only consider closed swarm instead of convoy and group pattern since closed swarm is more general [10].

A gradual trajectory pattern [6], denoted *rGpattern*, is designed to capture the gradual object moving trend. More precisely, a rGpattern is a maximal list of moving object clusters which satisfy the graduality constraint and integrity condition during at least min_t timestamps. The graduality constraint can be the increase or decrease of the number of objects and the integrity condition can be that all the objects should remain in the next cluster. A rGpattern can be defined as follows:

Definition 2 *rGpattern [6]. Given a list of clusters $C^* = c_1 \dots c_n$. C^* is a gradual trajectory pattern if:*

$$C^* = C^{\geq} \begin{cases} (1) : |C^*| \geq min_t. \\ \forall i \in \{1, \dots, n-1\}, \\ (2) : o(c_i) \subseteq o(c_{i+1}). \\ (3) : |c_n| > |c_1|. \\ (4) : \nexists c_m : C^* \cup c_m \textit{ is a } C^{\geq}. \end{cases} \qquad C^* = C^{\leq} \begin{cases} (1) : |C^*| \geq min_t. \\ \forall i \in \{1, \dots, n-1\}, \\ (2) : o(c_i) \supseteq o(c_{i+1}). \\ (3) : |c_n| < |c_1|. \\ (4) : \nexists c_m : C^* \cup c_m \textit{ is a } C^{\geq}. \end{cases}$$

Essentially, we have two kinds of rGpatterns, $rGpattern^{\geq}$ and $rGpattern^{\leq}$. For instance, see Figure 1, $rGpattern^{\geq} = c_1 c_4 c_6$ and $rGpattern^{\leq} = c_7 c_8$.

2.2 Problem Statement

Eliminating the number of uninteresting patterns is an emerging task in many real world cases. One of the proposed solutions is the MDL principle [4]. Let us start explaining this principle in the following definition:

Definition 3 *(Hypothesis). A hypothesis $\mathcal{P}$ is a set of patterns $\mathcal{P} = \{p_1, p_2, \ldots, p_h\}$.*

Given a scheme S, let $L_S(P)$ be the description length of hypothesis $\mathcal{P}$ and $L_S(O_{DB}|P)$ be the description length of data O_{DB} when encoded with the help of the hypothesis and an encoding scheme S. Informally, the MDL principle proposes that the best hypothesis always compresses the data most. Therefore, the principle suggests that we should look for hypothesis $\mathcal{P}$ and the encoding scheme S such that $L_S(O_{DB}) = L_S(\mathcal{P}) + L_S(O_{DB}|\mathcal{P})$ is minimized. For clarity sake, we will omit S when the encoding scheme is clear from the context. Additionally, the description length of O_{DB} given $\mathcal{P}$ is denoted as $L_{\mathcal{P}}(O_{DB}) = L(\mathcal{P}) + L(O_{DB}|\mathcal{P})$.

In this paper, the hypothesis is considered as a dictionary of movement patterns $\mathcal{P}$. Furthermore, as in [9], we assume that any number or character in data has a fixed length bit representation which requires a unit memory cell. In our context, the description length of a dictionary $\mathcal{P}$ can be calculated as the total lengths of the patterns and the number of patterns (i.e. $L(\mathcal{P}) = \sum_{p\in\mathcal{P}} |p| + |\mathcal{P}|$). Furthermore, the length of the data O_{DB} when encoded with the help of dictionary $\mathcal{P}$ can be calculated as $L(O_{DB}|\mathcal{P}) = \sum_{o\in O_{DB}} |o|$.

The problem of finding compressing patterns can be formulated as follows:

Definition 4 *(Compressing Pattern Problem). Given a moving object database O_{DB}, a set of pattern candidates $F = \{p_1, p_2, \ldots, p_m\}$. Discover an optimal dictionary $\mathcal{P}^*$ which contains at most K movement patterns so that:*

$$\mathcal{P}^* = \arg\min_{\mathcal{P}} \left(L^*_{\mathcal{P}}(O_{DB})\right) = \arg\min_{\mathcal{P}} \left(L^*(\mathcal{P}) + L^*(O_{DB}|\mathcal{P})\right), \mathcal{P}^* \subseteq F \quad (2)$$

A key issue in designing an MDL-based algorithm is: how can we encode data given a dictionary? The fact is that if we consider closed swarms individually, Krimp algorithm can be easily adapted to extract compression patterns. However, the issue here is that we have different patterns (i.e. closed swarms and rGpatterns) and Krimp algorithm has not been designed to deal with rGpatterns. It does not supply multi-pattern types in the dictionary that may lead to losing interesting ones. Furthermore, as mentioned before, we also have to address the pattern overlapping issue. In this work, we propose a novel overlapping allowed multi-pattern structures encoding scheme for moving object data.

3 Encoding Scheme

3.1 Movement Pattern Dictionary-based Encoding

Before discussing our encoding for moving object data, we revisit the encoding scheme used in the Krimp algorithm [14]. An itemset I is encoded with the help of itemset patterns by replacing every non-overlapping instance of a pattern occurring in I with a pointer to the pattern in a code table (dictionary). In this way, an itemset can be encoded to a more compact representation and decoded back to the original itemset.

Table 1. An illustrative example of database and dictionary in Figure 1. $\bar{0}$, $\bar{1}$ and $\bar{2}$ respectively are pattern types: closed swarm, $rGpattern^{\geq}$ and $rGpattern^{\leq}$.

O_{DB}	Encoded O_{DB}	Dictionary $\mathcal{P}$
$o_1 = c_1c_4c_6c_7c_8$	$o_1 = [p_1, 0][p_3, 1]$	
$o_2 = c_3c_4c_6c_7c_{10}$	$o_2 = c_3[p_1, 1][p_3, 0]c_{10}$	$p_1 = c_1c_4c_6, \bar{1}$
$o_3 = c_6$	$o_3 = [p_1, 2]$	$p_2 = c_2c_5c_7c_9, \bar{0}$
$o_4 = c_2c_5c_7c_9$	$o_4 = p_2$	$p_3 = c_7c_8, \bar{2}$
$o_5 = c_2c_5c_7c_9$	$o_5 = p_2$	

In this paper we use a similar dictionary-based encoding scheme for moving object database. Given a dictionary consisting of movement patterns $\mathcal{P} = \{p_1, \ldots, p_m\}$, an object $o \in O_{DB}$ containing a list of clusters is encoded by replacing instances of any pattern p_i in o with pointers to the dictionary. An important difference between itemset data and moving object data is that there are different kinds of movement patterns which have their own characteristic. The fact is that if a closed swarm cs occurs in an object o then all the clusters in cs are involved in o. While an object can involve in only a part of a rGpattern and viceversa.

For instance, see Figure 1, we can consider that o_2 joins the $rGpattern^{\geq} = c_1c_4c_6$ at c_4c_6. While, the closed swarm $cs = c_2c_5c_7c_9$ occurs in o_4 and o_5 entirely.

Property 1. (Encoding Properties). Given an object o which contains a list of clusters and a pattern $p = c_1 \ldots c_n$. p occurs in o or o contributes to p if:

$$\begin{cases} (1) : p \textit{ is a } rGpattern^{\geq}, \exists i \in [1, n] \big| \forall j \geq i, c_j \in o. \\ (2) : p \textit{ is a } rGpattern^{\leq}, \exists i \in [1, n] \big| \forall j \leq i, c_j \in o. \\ (3) : p \textit{ is a closed swarm}, \forall j \in [1, n], c_j \in o. \end{cases} \tag{3}$$

Proof. Case (1): after construction we have $o(c_i) \subseteq o(c_{i+1}) \subseteq \ldots \subseteq o(c_n)$. Additionally, $o \in o(c_i)$. Consequently, $o \in o(c_{i+1}), \ldots, o(c_n)$ and therefore $\forall j \geq i, c_j \in o$. Furthermore, in Case (2): we have $o(c_1) \supseteq o(c_2) \supseteq \ldots \supseteq o(c_{i-1})$. Additionally, $o \in o(c_{i-1})$. Consequently, $o \in o(c_1), \ldots, o(c_{i-1})$ and therefore $\forall j \leq i, c_j \in o$. In Case (3), we have $o \in O(cs) = \bigcap_{i=1}^{n} c_i$ and therefore $\forall j \in [1, n], c_j \in o$.

For instance, see Table 1, we can see that for each pattern, we need to store an extra bit to indicate the pattern type. Regarding to closed swarm, by applying Property 1, in the object o we only need to replace all the clusters, which are included in closed swarm, by a pointer to the closed swarm in the dictionary. However, in gradual trajectories (i.e. $rGpattern^{\geq}$, $rGpattern^{\leq}$), we need to store with the pointer an additional index to indicate the cluster c_i. Essentially, c_i plays the role of a starting involving point (resp. ending involving point) of the object o in a $rGpattern^{\geq}$ (resp. $rGpattern^{\leq}$).

As an example, consider dictionary $\mathcal{P}$ in Table 1. Using $\mathcal{P}$, o_1 can be encoded as $o_1 = [p_1, 0][p_3, 1]$ where 0 (in $[p_1, 0]$) indicates the cluster at index 0 in p_1, (i.e. c_1) and 1 (in $[p_3, 1]$) indicates the cluster at index 1 in p_3, i.e. c_8. While, o_4 can be encoded as $o_4 = p_2$, i.e. p_2 is a closed swarm.

3.2 Overlapping Movement Pattern Encoding

Until now, we have already presented the encoding function for different patterns when encoding an object o given a pattern p. In this section, the encoding scheme will be completed by addressing the pattern overlapping problem so that overlapped patterns can exist in the dictionary $\mathcal{P}$.

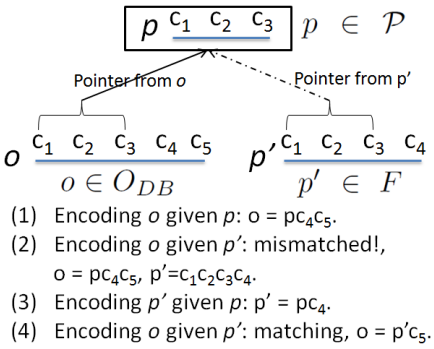


Fig. 5. An example of the approach.

See Figure 5, a selected pattern $p \in \mathcal{P}$ and a candidate $p' \in F$ overlap each other at $c_1c_2c_3$ on object o. Assume that o is encoded given p then $o = pc_4c_5$. As in Krimp algorithm, p' is still remained as origin and then p' cannot be used to encode o despite of p' occurs in o. This is because they are mismatched (i.e. $o = pc_4c_5, p' = c_1c_2c_3c_4$). To solve the problem, we propose to encode p' given p so that o and p' will contain the same pointer to p (i.e. $p' = pc_4$). Now, the regular encoding scheme can be applied to encode o given p' (i.e. $o = p'c_5$). We can consider that p and p' are overlapping but both of them can be included in the dictionary $\mathcal{P}$. **Note:** in our context, overlapped clusters are counted only once.

Main idea. Given a dictionary $\mathcal{P}$ and a chosen pattern p (i.e. will be added into $\mathcal{P}$), a set of pattern candidates F. The main idea is that we first encode the database O_{DB} given pattern p. Secondarily, we propose to encode all candidates $p' \in F$ given p in order to indicate the overlapping clusters between p and p'. After that, there are two kinds of pattern candidates which are encoded candidates and non-encoded candidates. Next, the best candidate in F will be put into $\mathcal{P}$ and used to encode O_{DB} and F. The process will be repeat until obtaining top-K patterns in the dictionary $\mathcal{P}$.

Let us consider the correlations between a pattern $p \in \mathcal{P}$ and a candidate $p' \in F$ to identify whenever encoding p' given p is needed. The correlation between p and p' is illustrated in Table 2. First of all, we do not allow overlap between two patterns of the same kind since they represent the same knowledge that may lead to extracting redundant information.

Table 2. Correlations between pattern p and pattern p' in F. O, Δ and X respectively mean *"overlapping allowed, regular encoding"*, *"overlapping allowed, no encoding"* and *"overlapping not allowed"*.

		p		
		cs	$rGpattern^{\geq}$	$rGpattern^{\leq}$
p'	cs	X	O	O
	$rGpattern^{\geq}$	Δ	X	O
	$rGpattern^{\leq}$	Δ	O	X

Next, if p is a closed swarm then p' do not need to be encoded given p. This is because there are objects which contribute to gradual trajectories p' but not closed swarm. These objects cannot be encoded using p and therefore p' needs to be remained the same and the regular encoding scheme can be applied. Otherwise, p' will never be chosen later since there are no objects in O_{DB} which match p'. For instance, see Figure 2, the objects o_1 and o_4 do not contribute to the closed swarm p. Thus, if the gradual trajectory p' is encoded given p to indicate the overlapping clusters $c_5c_6c_7$ then that leads to a mismatched statement between o_1, o_4 and the gradual trajectory p'.

Until now, we already have two kinds of candidates $p' \in F$ (i.e. non-encoded and encoded candidates). Next, some candidates will be used to encode the database O_{DB}. To encode an object $o \in O_{DB}$ given a non-encoded candidate p', the regular encoding scheme mentioned in Section 3.1 can be applied. However, given an encoded candidate

p', we need to perform an additional step before so that the encoding scheme can be applied regularly. This is because the two pointers referring to the same pattern $p \in \mathcal{P}$ from o (e.g. $[p,k]$) and from p' (e.g. $[p,l]$) can be different (i.e. $k \neq l$) despite the fact that p' is essentially included in o. That leads to a mismatched statement between o and p' and thus o cannot be encoded given p'.

For instance, see Figure 2, given a gradual trajectory pattern $rGpattern^{\geq}$ $p = c_3c_4c_5c_6c_7$, a closed swarm $p' = c_1c_2c_5c_6c_7c_9c_{10}$, the object $o_3 = c_1c_2c_4c_5c_6c_7c_9c_{10}$. We first encodes o_3 given p such that $o_3 = c_1c_2[p,1]c_9c_{10}$. Then, p' is encoded given p, i.e. $p' = c_1c_2[p,2]c_9c_{10}$. We can consider that the two pointers referring to p from o (i.e. $[p,1]$) and from p' (i.e. $[p,2]$) are different and thus o_3 and p' are mismatched. Therefore, o cannot be encoded given p' despite the fact that p' essentially occurs in o.

To deal with this issue, we simply recover uncommon clusters between the two pointers. For instance, to encode o_3 by using p', we first recover uncommon cluster such that $o_3 = c_1c_2c_4[p,2]c_9c_{10}$. Note that $[p,1] = c_4[p,2]$. Since $p' = c_1c_2[p,2]c_9c_{10}$, o_3 is encoded given p' such that $o_3 = p'c_4$.

Definition 5 *(Uncommon Clusters for $rGpattern^{\geq}$). Given a $rGpattern^{\geq}$, $p = c_1 \dots c_n$ and two pointers refer to p, $[p,k]$ and $[p,l]$ with $k \leq l$. $uncom(p,k,l) = c_kc_{k+1} \dots c_{l-1}$ is called an uncommon list of clusters between $[p,k]$ and $[p,l]$. Note that $[p,k] = c_kc_{k+1} \dots c_{l-1}[p,l]$.*

Similarly, we also have $uncom(p,k,l)$ in the case p is a $rGpattern^{\leq}$. Until now, we are able to recover uncommon clusters between two pointers which refer to a pattern. Now, we start proving that given an object $o \in O_{DB}$ and a candidate $p' \in F$, if p' occurs in o then o can be encoded using p' even though they contain many pointers to other patterns. First, let us consider if p is a $rGpattern^{\geq}$ and p' is a closed swarm.

Lemma 1. *Given a $rGpattern^{\geq}$, $p = c_1 \dots c_n$, an object o and a closed swarm $p' \in F$. In general, if o and p' refer to p then $o = x_o[p,k]y_o$ and $p' = x_{p'}[p,l]y_{p'}$. Note that $x_o, y_o, x_{p'}$ and $y_{p'}$ are lists of clusters. If o contributes to p' then:*

$$k \leq l \wedge o = x_o\ uncom(p,k,l)[p,l]\ y_o \tag{4}$$

Proof. After construction if $k > l$ then $\exists c_i \in \{c_l, \dots, c_k\} (\subseteq p)$ s.t. $c_i \in p' \wedge c_i \notin o$. Therefore, o does not contribute to p' (Property 1). That suffers the assumption and thus we have $k \leq l$. Deal to the Definition 5, $[p,k] = uncom(p,k,l)[p,l]$. Consequently, we have $o = x_o\ uncom(p,k,l)[p,l]\ y_o$.

By applying Lemma 1, we have $o = x_o\ uncom(p,k,l)[p,l]\ y_o$ and $p' = x_{p'}[p,l]y_{p'}$. Then we can apply the regular encoding scheme to encode o given p'. let us assume that each object $o \in O_{p'}$ has a common list of pointers to other patterns as $\overrightarrow{(p',o)} = \{([p_1,l_1],[p_1,k_1]), \dots, ([p_n,l_n],[p_n,k_n])\}$ where $\forall i \in [1,n]$, $[p_i,l_i]$ is the pointer from p' to p_i and $[p_i,k_i]$ is the pointer from o to p_i. If we respectively apply Lemma 1 on each pointer in $\overrightarrow{(p',o)}$ then o can be encoded given p'. Similarly, we also have the other lemmas for other pattern types.

Data description length computation. Until now, we have defined an encoding scheme for movement patterns. The description length of the dictionary in Table 1 is calculated as $L(\mathcal{P}) = |p_1|+1+|p_2|+1+|p_3|+1+|\mathcal{P}| = 3+1+4+1+2+1+2 = 14$. Similarly, description length of o_2 is $L(o_2|\mathcal{P}) = 1 + |[p_1,1]| + |[p_3,0]| + 1 = 6$.

Note: for each pattern, we need to consider an extra memory cell of pattern type. Additionally, for any given dictionary $\mathcal{P}$ and the data O_{DB}, the cost of storing the timestamp for each cluster is always constant regardless the size of the dictionary.

4 Mining Compression Object Movement Patterns

In this section we will present the two greedy algorithms which have been designed to extract a set of top-K movement patterns that compress the data best.

4.1 Naive Greedy Approach

Algorithm 1: NaiveCompo

Input : Database O_{DB}, set of patterns F, int K
Output: Compressing patterns $\mathcal{P}$
Input : Database O_{DB}, set of patterns F, int K
Output: Compressing patterns $\mathcal{P}$
begin
  $\mathcal{P} \longleftarrow \emptyset$;
  while $|\mathcal{P}| < K$ **do**
    foreach $p \in F$ **do**
      $O^d_{DB} \longleftarrow O_{DB}$;
      $L^*(O^d_{DB}|p) \longleftarrow CompressionSize(O^d_{DB}, p)$;
    $p^* \longleftarrow \arg\min_p L^*(O^d_{DB}|p)$;
    $\mathcal{P} \longleftarrow p^*$; $F \longleftarrow F \setminus \{p^*\}$;
    Replace all instances of p^* in O_{DB} by its pointers;
    Replace all instances of p^* in F by its pointers;
  output $\mathcal{P}$;
CompressionSize(O^d_{DB}, p)
begin
  $size \longleftarrow 0$;
  foreach $o \in O_{DB}$ **do**
    if $p.involved(o) = true$ **then**
      Replace instance of p in o by its pointers;
  foreach $o \in O_{DB}$ **do**
    $size \longleftarrow size + |o|$;
  $size \longleftarrow size + |p| + 1$;
  output $size$;

The greedy approach takes as input a database O_{DB}, a candidate set F and a parameter K. The result is the optimal dictionary which encodes O_{DB} best. Now, at each iteration of *NaiveCompo*, we select candidate p' which compresses the database best. Next, p' will be added into the dictionary $\mathcal{P}$ and then the database O_{DB} and F will be encoded given p'. The process is repeated until we obtain K patterns in the dictionary.

To select the best candidate, we generate a duplication of the database O^d_{DB} and for each candidate $p' \in F$, we compress O^d_{DB}. The candidate p' which returns the smallest data description length will be considered as the best candidate. Note that $p' = argmin_{p^* \in F}\big(L_{p^*}(O_{DB})\big)$. The NAIVECOMPO is presented in Algorithm 1.

4.2 Smart Greedy Approach

The disadvantage of naive greedy algorithm is that we need to compress the duplicated database O^d_{DB} for each pattern candidate at each iteration. However, we can avoid this computation by considering some useful properties as follows.

Given a pattern p', $\overline{O}_{p'}$ and $O_{p'}$ respectively are the set of objects that do not contribute to p' and the set of objects involving in p'. The compression gain which is the number of memory cells we earned when adding p' into dictionary can be defined as $gain(p', \mathcal{P}) = L_{\mathcal{P}}(O_{DB}) - L_{\mathcal{P} \cup p'}(O_{DB})$.

The fact is that we can compute the compression gain by scanning objects $o \in O_{p'}$ with p'. Each pattern type has its own compression gain computation function. Let us start presenting the process by proposing the property for a closed swarm p'.

Property 2. Given a dictionary $\mathcal{P}$, a closed swarm $p' \in F$. $gain(p', \mathcal{P})$ is computed as:

$$gain(p', \mathcal{P}) = |O_{p'}| \times |p'| - \Big(\sum_{o}^{O_{p'}} \sum_{i}^{\overrightarrow{(p', o)}} |l_i - k_i| + |p'| + |O_{p'}| + 2\Big) \qquad (5)$$

Proof. After construction we have $L_{\mathcal{P}\cup p'}(O_{DB}) = L(\mathcal{P} \cup p') + L(O_{DB}|\mathcal{P} \cup p') = (L(\mathcal{P}) + |p'| + 2) + L(\overline{O}_{p'}|\mathcal{P}) + L(O_{p'}|\mathcal{P} \cup p')$. Note that $L(\overline{O}_{p'}|\mathcal{P}) = L(\overline{O}_{p'}|\mathcal{P} \cup p')$. Furthermore, $\forall o \in O_{p'} : L(o|\mathcal{P} \cup p') = L(o|\mathcal{P}) - |p'| + 1 + \sum_i^{\overrightarrow{(p',o)}} |l_i - k_i|$. Thus, $L(O_{p'}|\mathcal{P}\cup p') = \sum_{o\in O_{p'}} L(o|P\cup p') = L(O_{p'}|\mathcal{P}) - |O_{p'}| \times |p'| + \sum_o^{O_{p'}} \sum_i^{\overrightarrow{(p',o)}} |l_i - k_i| + |O_{p'}|$. Therefore, we have $L_{\mathcal{P}\cup p'}(O_{DB}) = L(\mathcal{P}) + L(\overline{O}_{p'}|\mathcal{P}) + L(O_{p'}|\mathcal{P}) - |O_{p'}| \times |p'| + \big(\sum_o^{O_{p'}} \sum_i^{\overrightarrow{(p',o)}} |l_i - k_i| + |p'| + |O_{p'}| + 2\big)$. Note that $L(O_{DB}|\mathcal{P}) = L(\overline{O}_{p'}|\mathcal{P}) + L(O_{p'}|\mathcal{P})$. Consequently, we have $gain(p', \mathcal{P}) = |O_{p'}| \times |p'| - \big(\sum_o^{O_{p'}} \sum_i^{\overrightarrow{(p',o)}} |l_i - k_i| + |p'| + |O_{p'}| + 2\big)$.

By applying Property 2, we can compute the compression gain when adding a new closed swarm p' into the dictionary $\mathcal{P}$. In the Equation 5, the compression $gain(p', \mathcal{P})$ depends on the size of p', $O(p')$ and the number of uncommon clusters that can be computed by scanning p' with objects $o \in O(p')$ without encoding O_{DB}. Due to the space limitation, we will not describe properties and proofs for the other pattern types (i.e. $rGpattern^{\geq}, rGpattern^{\leq}$) but they can be easily derived in a same way as Property 2.

To select the best candidate at each iteration, we need to chose the candidate which returns the best compression gain. SMARTCOMPO is presented in the Algorithm 2.

5 Experimental Results

Algorithm 2: SmartCompo

```
Input  : Database O_DB, set of patterns F, int K
Output: Compressing patterns P
1  begin
2      P ⟵ ∅;
3      while |P| < K do
4          foreach p ∈ F do
5              L*(O_DB|p) ⟵ Benefit(O_DB, p);
6          p* ⟵ arg min_p L*(O_DB|p);
7          P ⟵ p*; F ⟵ F \ {p*};
8          Replace all instances of p* in O_DB by its pointers;
9          Replace all instances of p* in F by its pointers;
10     output P;
11 Benefit(O^d_DB, p)
12 begin
13     b ⟵ 0;
14     foreach o ∈ O_DB do
15         if p.involved(o) = true then
16             b ⟵ b + benefit(o, p);
17     b ⟵ b + |p| + 1;
18     output b;
```

A comprehensive performance study has been conducted on real-life datasets. All the algorithms are implemented in C++, and all the experiments are carried out on a 2.8GHz Intel Core i7 system with 4GB Memory. The system runs Ubuntu 11.10 and g++ 4.6.1.

As in [10] [6], the two following datasets[3] have been used during experiments: Swainsoni dataset includes 43 objects evolving over 764 different timestamps. The dataset was generated from July 1995 to June 1998. Buffalo dataset concerns 165 buffaloes and the tracking time from year 2000 to year 2006. The original data has 26610 reported locations and 3001 timestamps. Similarly to [7] [10], we first use linear interpolation to fill in the missing data. Furthermore, DBScan [2] ($MinPts = 2; Eps = 0.001$) is applied to generate clusters at each timestamp. In the comparison, we compare the set of patterns produced by SmartCompo with the set of closed swarms extracted by *ObjectGrowth* [10] and the set of gradual trajectories extracted by *ClusterGrowth* [6].

[3] http://www.movebank.org

(a) $rGpattern \geq$ (b) Closed swarm (c) $rGpattern^{\leq}$

Fig. 6. Top-3 typical compression patterns.

Effectiveness. We compare the top-5 highest support closed swarms, the top-5 highest covered area gradual trajectory patterns and the top-5 compression patterns from Swainsoni dataset. Each color represents a Swainsoni trajectory involved in the pattern.

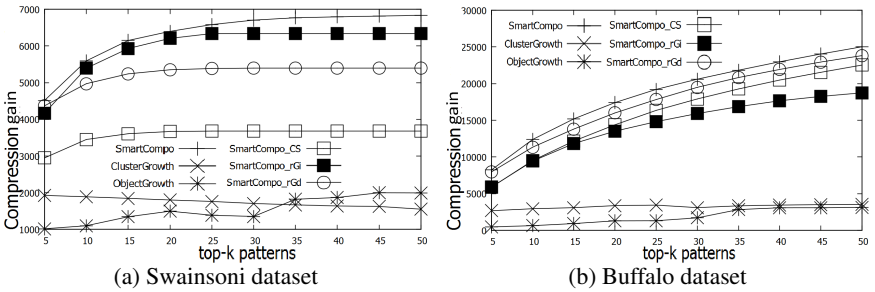
Top-5 closed swarms are very redundant since they only express that Swainsonies move together from North America to Argentina. Similarly, top-5 rGpatterns are also redundant. They express the same knowledge that is *"from 1996-10-01 to 1996-10-25, the more time passes, the more objects are following the trajectory* $\{Oregon\rangle\ Nevada\rangle\ Utah\rangle\ Arizona\rangle\ Mexico\rangle\ Colombia\}$*"*.

Figure 6 illustrates 3 patterns among 5 extracted ones by using SmartCompo. The $rGpattern^{\geq}$ expresses the same knowledge with the mentioned rGpattern in the top highest covered area. The closed swarm expresses new information that is *"after arriving South America, the Swainsonies tend to move together to Argentina even some of them can leave their group"*. Next, the $rGpattern^{\leq}$ shows that *"the Swainsonies return back together to North America from Argentina (i.e. 25 objects at Argentina) and they will step by step leave their group after arriving Guatemala (i.e. 20 objects at Guatemala) since they are only 2 objects at the last stop, i.e. Oregon State"*.

Compressibility. We measure the compressibility of the algorithms by using their top-K patterns as dictionaries for encoding the data. Since NaiveCompo and SmartCompo provides the same results, we only show the compression gain of SmartCompo.

Regarding to SmartCompo, the compression gain could be calculated as the sum of the compression gain returned after each greedy step with all kinds of patterns in F. For each individual pattern type, compression gain is calculated according to the greedy encoding scheme used for SmartCompo. They are respectively denoted as $SmartCompo\_CS$ (i.e. for closed swarms), $SmartCompo\_rGi$ (i.e. for $rGpattern^{\geq}$) and $SmartCompo\_rGd$ (i.e. for $rGpattern^{\leq}$). Additionally, to illustrate the difference between MDL-based approaches and standard support-based approaches, we also employ the set of top-K highest support closed swarms and top-K highest covered area gradual trajectories patterns.

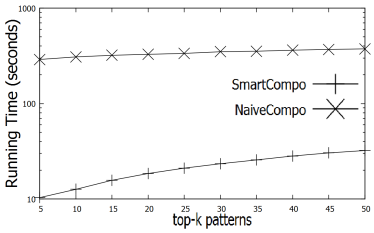
Figure 7 shows the compression gain of different algorithms. We can consider that top-K highest support or covered area patterns cannot provide good compression gain since they are very redundant. Furthermore, if we only consider one pattern type, we cannot compress the data best since the compression gains of SmartCompo_CS, SmartCompo_rGi and SmartCompo_rGd are always lower than SmartCompo. This is because the pattern distribution in the data is complex and different patterns can cover different parts of the data. Thus, considering one kind of patterns results in losing interesting pat-



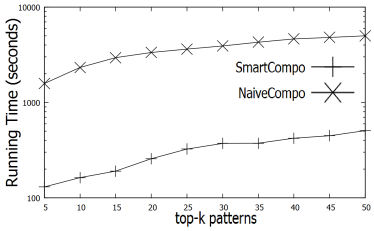
(a) Swainsoni dataset (b) Buffalo dataset

Fig. 7. Compressibility (higher is better) of different algorithms.

terns and not good compression gain. By proposing overlapping allowed multi-pattern structure encoding scheme, we are able to extract more informative patterns.



(a) Swainsoni dataset



(b) Buffalo dataset

Fig. 8. Running time.

One of the most interesting phenomena is that the Swainsonies and Buffaloes have quite different movement behavior. See Figure 7a, we can consider that $rGpattern^{\geq}$ is the most representative movement behavior of Swainsonies since they compress the data better than the two other ones. While closed swarm is not as representative as the other patterns. This is because it is very easy for Swainsonies which are birds to leave the group and congregate again at later timestamps. However, this movement behavior is not really true for Buffaloes. See Figure 7b, it clear that the compression gains of closed swarms, $rGpattern^{\geq}$ and $rGpattern^{\leq}$ have changed. The three kinds of patterns have more similar compression gain than the ones in Swainsonies. It means that Buffaloes are more closed to each other and they move in a dense group. Thus closed swarm is more representative compare to itself in Swainsoni dataset. Furthermore, the number of Buffaloes is very difficult to increase in a group and thus SmartCompo_rGi is lower than the two other ones.

Running Time. In our best knowledge, there are no previous work which address mining compression movement pattern issue. Thus, we only compare the two proposed approaches in order to highlight the differences between them. Running time of each algorithm is measured by repeating the experiment in compression gain experiment.

As expected, SmartCompo is much faster than NaiveCompo (i.e. Figure 8). By exploiting the properties, we can directly select the best candidate at each iteration. Consequently, the process efficiency is speed up.

6 Related Work

Mining informative patterns can be classified into 3 main lines: MDL-based approaches, statistical approaches based on hypothesis tests and information theoretic approaches.

The idea of using data compression for data mining was first proposed by R. Cilibrasi et al. [1] for data clustering problem. This idea was also explored by Keogh et

al. [8], who propose to use compressibility as a measure of distance between two sequences. In the second research line, the significance of patterns is tested by using a standard statistical hypothesis assuming that the data follows the null hypothesis. If a pattern pass the test it is considered significant and interesting. For instance, A. Gionis et al. [3] use swap randomization to generate random transactional data from the original data. A similar method is proposed for graph data by R. Milo et al. [11]. Another research direction looks for interesting sets of patterns that compress the given data most (i.e. MDL principle). Examples of this direction include the Krimp algorithm [14] and Slim algorithm [12] for itemset data and the algorithms for sequence data [9].

7 Conclusion

We have explored an MDL-based strategy to compress moving object data in order to: 1) select informative patterns, 2) combine different kinds of movement patterns with overlapping allowed. We supplied two algorithms NaiveCompo and SmartCompo. The latter one exploits smart properties to speed up the whole process obtaining the same results to the naive one. Evaluations on real-life datasets show that the proposed approaches are able to compress data better than considering just one kind of patterns.

References

1. R. Cilibrasi and P. M. B. Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005.
2. M. Ester, H. P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.
3. A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. *TKDD*, 1(3), 2007.
4. P. Grunwald. The minimum description length principle. *The MIT Press*, 2007.
5. J. Gudmundsson and M. van Kreveld. Computing longest duration flocks in trajectory data. In *ACM GIS 06*, 2006.
6. P. N. Hai, D. Ienco, P. Poncelet, and M. Teisseire. Ming time relaxed gradual moving object clusters. In *ACM SIGSPATIAL GIS*, 2012.
7. H. Jeung, M. L. Yiu, X. Zhou, C. S. Jensen, and H. T. Shen. Discovery of convoys in trajectory databases. *Proc. VLDB Endow.*, 1(1):1068–1080, August 2008.
8. E. J. Keogh, S. Lonardi, C. A. Ratanamahatana, L. Wei, S. H. Lee, and J. Handley. Compression-based data mining of sequential data. *DMKD.*, 14(1):99–129, 2007.
9. H. T. Lam, F. Moerchen, D. Fradkin, and T. Calders. Mining compressing sequential patterns. In *SDM*, pages 319–330, 2012.
10. Z. Li, B. Ding, J. Han, and R. Kays. Swarm: mining relaxed temporal moving object clusters. *Proc. VLDB Endow.*, 3(1-2):723–734, September 2010.
11. R. Milo, S. Shen-Orr, S. Itzkovits, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594), 2002.
12. K. Smets and J. Vreeken. Slim: Directly mining descriptive patterns. In *SDM*, 2012.
13. L. A. Tang, Y. Zheng, J. Yuan, J. Han, A. Leung, C.-C. Hung, and W.-C. Peng. On discovery of traveling companions from streaming trajectories. In *ICDE*, pages 186–197, 2012.
14. J. Vreeken, M. Leeuwen, and A. Siebes. Krimp: Mining itemsets that compress. *Data Min. Knowl. Discov.*, 23(1):169–214, 2011.
15. Y. Wang, E. P. Lim, and S. Y. Hwang. Efficient mining of group patterns from user movement data. *Data Knowl. Eng.*, 57(3):240–282, 2006.
16. K. Zheng, Y. Zheng, J. Yuan, and S. Shang. On discovery of gathering patterns from trajectories. In *ICDE*, 2013.

Appendix B

Data Stream Classification: Active Learning

JMLR: Workshop and Conference Proceedings 29:1–16, 2014 BIGMINE 2014

High density-focused uncertainty sampling for active learning over evolving stream data

Dino Ienco DINO.IENCO@TELEDETECTION.FR
UMR TETIS, Irstea, Montpellier, France
LIRMM, Montpellier, France

Bernhard Pfahringer BERNHARD@CS.WAIKATO.AC.NZ
University of Waikato, Hamilton, New Zealand

Indrė Žliobaitė INDRE.ZLIOBAITE@AALTO.FI
Aalto University, Espoo, Finland
Helsinki Institute for Information Technology, Espoo, Finland

Abstract

Data labeling is an expensive and time-consuming task, hence carefully choosing which labels to use for training a model is becoming increasingly important. In the active learning setting, a classifier is trained by querying labels from a small representative fraction of data. While many approaches exist for non-streaming scenarios, few works consider the challenges of the data stream setting. We propose a new active learning method for evolving data streams based on a combination of density and prediction uncertainty (DBALStream). Our approach decides to label an instance or not, considering whether it lies in an high density partition of the data space. This allows focusing labelling efforts in the instance space where more data is concentrated; hence, the benefits of learning a more accurate classifier are expected to be higher. Instance density is approximated in an online manner by a sliding window mechanism, a standard technique for data streams. We compare our method with state-of-the-art active learning strategies over benchmark datasets. The experimental analysis demonstrates good predictive performance of the new approach.

Keywords: Active learning, Data streams, Density-based clustering

1. Introduction

Today, more data are being generated continuously than ever before. Streaming data pose serious challenges to data analysis researchers and practitioners. For learning predictive models on streaming data one needs to have continuous access to the true values of the target variable (the true class labels). This labeling phase is usually an expensive and tedious task for human experts. Consider, for example, textual news arriving as a data stream. The goal is to predict if a news item will interesting to given user at a given time. The interests and preferences of the user may change over time. To obtain training data, news items need to be labeled as interesting or not interesting. This requires human labor, which is time consuming and costly. For instance, Amazon Mechanical Turk[1] offers a marketplace for intelligent human labeling.

1. https://www.mturk.com

Labeling can also be costly because it may require expensive, intrusive or destructive laboratory tests. Consider a production process in a chemical plant where the goal is to predict the quality of production output. The relationship between input and output quality might change over time due to constant manual tuning, complementary ingredients or replacement of physical sensors. In order to know the quality of the output (the true label) a laboratory test needs to be performed, which is costly. Under such conditions it may be unreasonable to require true labels for all incoming instances.

A way to alleviate this issue is to query labels for a small representative portion of data. The main challenge is how to select a good subset of instances for learning a model. Such a learning scenario is referred to as *active learning* (Settles, 2010; Fu et al., 2013).

Active learning studies how to label selectively instead of asking for all true labels. It has been extensively studied in pool-based (Lewis and Gale, 1994) and online settings (Cohn et al., 1994). In pool-based settings the decision concerning which instances to label is made by ranking all historical instances (e.g. according to uncertainty) while in online active learning each incoming instance is compared to an uncertainty threshold and the system asks for the true label if the threshold is exceeded. The main difference between online active learning and active learning in data streams is in expectations around changes. In data streams the relationship between the input data and the label may change (concept drift) and these changes can happen anywhere in the instance space while online learning assumes a stationary relationship between examples and their labels. In order to capture concept drifts, which can happen anywhere in the data, it is important to characterize and understand how the examples are related to each other in the data space.

As a recent work in the pool-based settings suggests (Fu et al., 2013), considering the density around instances can improve the results. In the more dense regions in the instance space the benefits of updating the classifier with new data are expected to be larger, because it is likely to improve future classification accuracy for more instances.

The concept of density around instance is usually exploited in unsupervised learning (such as clustering (Rodriguez and Laio, 2014; Tomasev et al., 2014)) where local density of a point can be used to initialize cluster centroids. Local density can be directly estimate as the number of points closer to the considered instance (Rodriguez and Laio, 2014) or derived indirectly (Tomasev et al., 2014) as the number of times the instance appears as a neighbor of other example. These approaches widely demonstrate their efficacy for clustering tasks.

Unfortunately, very few works for active learning in data stream integrate the density (Ienco et al., 2013) in order to choose suitable queries instance. Motivated by this fact we introduce a new density-focused uncertainty sampling method that uses an online density approximation to guide the selection of labeling candidates in the data stream. Our approach works in a fully incremental way: in order to estimate the density of the neighborhood of an incoming instance, we use a sliding window mechanism to compare it to a recent portion of the historical data.

The remainder of this paper is organized as follows. Section 2 overviews existing active learning approaches for data streams and makes connections with semi-supervised learning in data streams. The proposed methodology is presented in Section 3. In Section 4 we present experimental results for a number of real world datasets and we also supply a sensitivity analysis of the sliding window size. Finally, Section 5 concludes the study.

2. Related Work

Online active learning has been the subject of a number of studies, where the data distribution is assumed to be static (Helmbold and Panizza, 1997; Sculley, 2007; Cohn et al., 1994; Attenberg and Provost, 2011). The goal is to learn an accurate model incrementally, without assuming to have all training instances available at the beginning, with minimum labeling effort. One example of such a scenario is malicious URL detection where training data are arrive sequentially Zhao and Hoi (2013). In contrast, in the evolving data streams setting, which is the subject of our study, the goal is to continuously update a model over time so that accuracy is maintained as data distribution is changing. The problem of label availability in evolving data streams has been the subject of several recent studies (Klinkenberg, 2001; Widyantoro and Yen, 2005; Masud et al., 2011; Fan et al., 2004; Huang and Dong, 2007; Zliobaite et al., 2014; Ienco et al., 2013) that fall into three main groups.

The first group of works uses semi-supervised learning approaches to label some of the unlabeled data automatically (Klinkenberg, 2001; Widyantoro and Yen, 2005; Masud et al., 2011), which can only work under the assumption that the class conditional distribution does not change, or, in other words, they assume that their is no concept drift. Semi-supervised learning approaches are conceptually different from the active learning approaches, that are the subject of our study, since the former can only handle changes in the input data distribution, changes in the relation between the input data and the target label cannot be spotted without querying an external oracle as is done in active learning.

The second group of works process data in batches implicitly or explicitly assuming that data is stationary within batches (Lindstrom et al., 2010; Masud et al., 2010; Fan et al., 2004; Huang and Dong, 2007). Such approaches require an external mechanism to handle concept drift. Lindstrom et al. (2010) use uncertainty sampling to label the most representative instances within each new batch. They do not explicitly detect changes, instead they use a sliding window approach, which discards the oldest instances. Masud et al. Masud et al. (2010) use uncertainty sampling within a batch to request labels. In addition, they use the unlabeled instances with their predicted labels for training, making it also another semi-supervised learning approach. A few works integrate active learning and change detection (Fan et al., 2004; Huang and Dong, 2007) in the sense that they first detect change and only if change is detected do they ask for representative true labels using offline active learning strategies designed for stationary data. In this scenario drift handling and active learning can be considered as two mechanisms operating in parallel, but doing so independently. This is the main difference between this scenario and the last one, which combines the two mechanisms more closely together.

Finally, the third group of works use randomization to capture possible changes in the class conditional distribution (Zhu et al., 2007; Zliobaite et al., 2014; Cesa-Bianchi et al., 2006). Cesa-Bianchi et al. (2006) develop an online active learning method for a perceptron based on selective sampling using a variable labeling threshold $b/(b+|p|)$, where b is a parameter and p is the prediction of the perceptron. The threshold itself is based on certainty expectations, while the labels are queried at random. This mechanism could allow adaptation to changes, although they did not explicitly consider concept drift.

Differently from previous works, and in the same direction of our idea, Ienco et al. (2013) proposes a clustering-based active learning method for data streams. In that work density is

captured by means of clustering in a batch-incremental scenario. The clustering step allows to perform stratified sampling considering dense areas to selecting examples for labelling.

3. Method

In this section we describe our new approach DBALStream (**D**ensity **B**ased **A**ctive **L**earning for Data **Stream**s). We work in a fully incremental scenario in which each instance is processed as soon as it arrives. We can define a data stream as $S := \{x_1, x_2, ...x_n, ...\}$ where each x_i is a new instance arriving at time i. This scenario is general enough to model arbitrary real-world data streams. Given a data stream S and a budget b we want to learn a classifier cl with only b of the instances in the stream. How to select the instances to query is challenging for any active learning strategy.

The proposed strategy aims at modelling density around an instance, and is guided by the following hypothesis: an instance is more important to label if it lies in a dense area. For example, suppose we are classifying the sentiments towards news items, arriving online. There are a lot of incoming news about the political situation in Ukraine, and very few news items about research in Antarctica. Suppose, sentiments towards Antarctica news are currently more uncertain. We can label the Antarctica news item and learn to classify in this context very accurately, but if similar news items arrive very rarely in the future, we have little gain. Perhaps, we would better label news related to Ukraine, which are frequent (instance space of high density). This way, assuming that there is no sudden change in density, we can expect improvement in future classification accuracy on more instances.

In order to implement this intuition, given a data point x_i we model its density as the number of times x_i is the nearest neighbor of other instances. We use this value as an estimate of the density around a particular instance and in particular, we use this measure to understand if an instance x_i lies in a dense area, or not.

In a batch scenario this operation can be performed over the whole dataset in order to obtain a good picture of the global behavior.

In our setting, as we deal with data streams, it is infeasible to store all data and perform such an operation on the whole data stream. Therefore, we only consider a window, or buffer, of previously processed examples in order to estimate a local density factor for a new incoming example. Given a window W of previous examples and a data structure $MinDist$ that stores the minimum measured distance for each of the instances in W, the local density factor (ldf) of a new instance x_i is defined as follows:

$$ldf(x_i) = \sum_{x_j \in W} \mathbb{I}\{MinDist(x_j) > dist(x_i, x_j)\} \quad (1)$$

where $\mathbb{I}$ is an indicator function that returns 1 if the condition is true, and 0 otherwise. The function $dist(\cdot, \cdot)$ is a distance function defined over two examples x_i and x_j of the stream. Procedurally speaking, the $MinDist$ data structure is updated each time the indicator function is verified. This means that if $MinDist(x_j) > dist(x_i, x_j)$ then $MinDist(x_j)$ is updated with the value $dist(x_i, x_j)$. The minimum distance computed between x_i and each element of W is used as initial value for $MinDist(x_i)$. Another important aspect to underline is that both W and $MinDist$ cannot exceed a predefined size and for this reason

they work as a queue data structure (First In First Out): when the limit is reached the first instance, the oldest one, in the queue is removed and the new one is pushed at the end.

The local density factor of an instance can supply information about its importance, but it is not enough to understand if an example might be useful or not to label. As shown in (Fu et al., 2013), an important factor to consider is the uncertainty related to an instance that is usually employed in general active learning approaches for data streams (Zliobaite et al., 2014).

In order to combine both these aspects, local density factor and uncertainty of an instance, we build our new proposal extending one of the frameworks proposed in (Zliobaite et al., 2014). We extend the framework used for the *Uncertainty Strategy with Randomization*. This method relies on a randomized dynamic threshold, trying to label the least certain instances within a time interval. This threshold adapts, depending on the incoming data, to align with the budget, and it is also adjusted by a normal random factor. If a classifier becomes more certain, the threshold is increased, so that only the most uncertain instances are captured for labelling, otherwise the threshold is decreased to extract more information for improving the classifier.

The original framework employs the maximum a posteriori probability of the classifier in order to quantify the uncertainty of an instance. This method prefers instances on which the current classifier is less confidence (Fu et al., 2013) and can be formally defined as follow:

$$Confidence(x_i) = max_{cl} P_L(y_{cl}|x_i)$$

A different approach to model the uncertainty of the instance x_i is to consider not only the maximum a posteriori probability, but considering also the second most probable class label. Thus this approach is Margin-based (Fu et al., 2013) and it is prone to select instances with minimum margin between posterior probabilities of the two most likely class labels. The margin is defined as:

$$Margin(x_i) = P_L(y_{cl_1}|x_i) - P_L(y_{cl_2}|x_i)$$

where the cl_1 and cl_2 are respectively the class with the maximum a posteriori probability and the second most probable class.

Our proposal is depicted in Algorithm 1. The algorithm takes as input the instance to evaluate x_t, the stream classifier L, the budget percentage b, the adjusting step s, the threshold θ, the window of previous collected instances W and the $MinDist$ data structure that stores the minimum distance for each element in W. The algorithm first computes the number of times the instance x_t is the nearest neighbor of an instance in W. This operation is performed using Formula 1 and implemented by the function $compute\#timesNN(x_t, W, MinDist)$. This procedure returns an integer representing the local density around example x_t and, at the same time, it updates W and $MinDist$ according to their maximum size. Next, the procedure tests two conditions: the first one ($u/t < b$) verifies the budget constraint, which disallow any new labelling when the budget is already exceeded; the second condition ($lrd(x_i) \neq 0$) takes into account the local density function. In particular this test does not allow to ask for labels for instances that lie outside any local dense areas of the data space. If either of the conditions fails, then the algorithm returns false, otherwise the margin for the instance x_t is computed.

Line 4 to 12 are inspired by the approach proposed in (Zliobaite et al., 2014) where the randomized adaptive threshold method was firstly introduced. In Line 4 the threshold θ is randomized employing a variable ϵ sampled from a Normal distribution $N(0,1)$. This randomization step allows to occasionally label instances that are far from the decision boundary. Line 5 checks the uncertainty of x_t w.r.t. the randomized threshold θ_{ran}. If the test is positive, then the instance will be used to train the classifier and the active learning strategy consumes some budget (Line 6). The threshold θ is updated, considering the adjusting step (Line 7) and the procedure terminates. If the test (Line 5) fails, the Algorithm 1 returns false, no budget is spent and the threshold θ is relaxed in order to increase the chance of labeling the next time.

The DBALStream algorithm is employed in the general data stream classification framework proposed in Section 3.1. In particular, it is used to decide if the true label y_t for the instance x_t should be asked for, or not.

Regarding the computational complexity of our method, the most time consuming steps are the computation of $ldf(x_i)$ and the updating of $MinDist$ structure. Both operations have linear complexity in the size of $|W|$ and they can be done at the same time (we need to scan once the window W for each incoming instance x_i).

Algorithm 1 DBALStream(x_t, L, b, s, u, θ, W, $MinDist$)

Require: x_t: the new instance in the data stream, L: learned classifier, b: budget, s: adjusting step, u: number of instances labeled till now, θ: the actual threshold value, W: set of previous instances used to computed the local density function, $MinDist$: minimum distances for each instance in the window W

1: $ldf(x_t) = compute\#timesNN(x_t, W, MinDist)$
2: **if** $(u/t < b$ and $ldf(x_t) \neq 0)$ **then**
3: $\quad margin(x_t) = P_L(y_{cl1}|x_t) - P_L(y_{cl2}|x_t)$
4: $\quad \theta_{ran} = \theta \times \epsilon$ where $\epsilon \in N(0,1)$
5: $\quad$ **if** $margin(x_t) < \theta_{ran}$ **then**
6: $\quad\quad u = u + 1$
7: $\quad\quad \theta = \theta * (1 - s)$
8: $\quad\quad$ **return true**
9: $\quad$ **else**
10: $\quad\quad \theta = \theta * (1 + s)$
11: $\quad\quad$ **return false**
12: $\quad$ **end if**
13: **end if**
14: **return false**

3.1. General Active Learning Classification Framework

Algorithm 2 presents the general active learning classification framework. It requires as input the data stream S and the budget percentage b. At the beginning (line 1 and 2) the framework initializes the different parameters. In particular we set the θ threshold equal to 1 and the step option s to 0.01. The step option s is used for increasing or decreasing the threshold in the DBALStream procedure.

We use DDDM change detection technique (Gama et al., 2004): the accuracy of the classifier is monitored during the streaming process. If the accuracy starts to decrease (*change warning is signaled* line 6-8) a new background classifier L_n is generated.

At line 9-11 both classifiers are trained in parallel with the incoming instance x_t. If a *change is detected*, the classifier L is replaced by L_n (line 13-15).

Classifier performances are evaluated by means of the prequential schema. The evaluation through the prequential setting involves two steps: i) classify an instance, and ii) use the same instance to train the learner. To implement this strategy, each new incoming instance x_t is first classified using L and only after this step the approach decides if a label is waned for training, or not. The accuracy is computed over the classification results of L. This process continues until the end of the data stream is reached.

Algorithm 2 Active Learning Classification Framework

Require: b: labeling budget
Require: S the data stream
1: $\theta = 1$, $u = 0$, $s = 0.01$, $W = \emptyset$, $MinDist = \emptyset$
2: L =classifier, L_n =classifier
3: **for** $x_t \in S$ **do**
4: **if** (DBALStream(x_t, L, b, s, u, θ, W, $MinDist$)) **then**
5: request the true label y_t of instance x_i
6: **if** change warning is signaled **then**
7: start a new classifier L_n
8: **end if**
9: train classifier L with (x_t,y_t)
10: **if** L_n exists **then**
11: train classifier L_n with (x_t,y_t)
12: **end if**
13: **if** change is detected **then**
14: replace classifier L with L_n
15: **end if**
16: **end if**
17: **end for**

4. Experiments

In this section we evaluate the performance and the quality of the proposed method DBALStream. We compare our algorithm with state of the art methods that are explicitly designed for active learning over data streams. We use the prequential evaluation procedure: each time an instance arrives, we first test the algorithm on it, and then we decide on whether to pay the cost for labeling it and subsequently use it as an input for updating the classifier. In order to evaluate the performance of the different strategies we employ classification accuracy. We use two methods proposed by (Zliobaite et al., 2014): *Random* and *Rand Unc*. The first one is *Random*. It randomly chooses examples for labeling. The second one (*Rand Unc*), proposed in (Zliobaite et al., 2014), uses a randomized variable uncertainty strategy that combines the randomization with maximum a posteriori probability and an adaptive method to avoid consuming too much of the budget when a consecutive run of easy instances is encountered. The last competitor *ACLStream* is a clustering-based active learning approach proposed by Ienco et al. (2013). This approach works in a batch-incremental scenario. It performs stratified sampling, where at first a clustering solution over a batch of data is obtained, and then instances are chosen for labelling considering a combination of their own uncertainty and the uncertainty of the cluster they belong to.

For all methods a warm-up period is introduced. The first 500 instances of each dataset are all labeled and used to train the initial model used by the specific approach. Evaluation using active learning only starts after this warm-up step.

All the methods need an internal classification algorithm to perform the classification and to produce the maximum a posteriori probability. For this reason for all the approaches we use the classifier proposed in (Gama et al., 2004). This classifier is able to adapt itself to drift situations: when the accuracy of the classifier begins to decrease, a new classifier is built and trained with new incoming instances. Considering the batch size, for *ACLStream*, following the original paper, we use a window size of 100 instances and the number of clusters is set to 5. As all the approaches are nondeterministic, each result, for each of the strategies, is averaged over 30 runs.

All the experiments are performed using the MOA data stream software suite (Bifet et al., 2010). MOA is an open source software framework implemented in Java designed specifically for data stream mining.

4.1. Datasets

To evaluate all the algorithms we use four real world benchmark datasets: *Electricity*, *CoverType*, *Airlines*, *KDD99*. *Electricity* data (Harries et al., 1998) is a popular benchmark in evaluating streaming classifiers. The task is to predict the rise or fall of electricity prices (demand) in New South Wales (Australia), given recent consumption and prices in the same and neighboring regions. *Cover Type* data (Bache and Lichman, 2013) is also often used as a benchmark for evaluating stream classifiers. The task is to predict forest cover types or types of vegetation from cartographic variables. Inspired by (Ikonomovska et al., 2010) an *Airlines* dataset was constructed using the raw data from US flight control. The task is to predict whether a given flight will be delayed, given the information of the scheduled departure. The last dataset, *KDD99*, is commonly used as a benchmark anomaly detection task but recently it has also been employed as a dataset for testing data stream algorithms (Masud et al., 2011). One of the big problems with this dataset is the big amount of redundancy among instances. To solve this problem we use the cleaned version named NSL-KDD[2]. To build the final dataset we join both training and test data. Table 1 presents summary characteristics of the datasets. This collection includes both binary and multi-class classification problems, datasets with different numbers of instances (varying between 42k to 829k) and different numbers of features (from 7 to 54). Even though *CoverType*, *KDD99* have no time order variable, we assume that they are presented in time order.

In order to better characterize this collection of datasets, we show for each of them, how the class variable evolves over time. For this purpose, we divided each dataset in batches and for each batch we plot its class values distribution. As the different datasets have different sizes, we choose an adapted batch size to allow for a clear trend visualization. This information can be useful to understand which dataset has dramatic (resp. smooth) changes considering the target variable. Logically, quick changes in the target variable requires more flexible classifiers than stationary situations. The characterization of the datasets is presented in Figure 1. The X axis represents the batch number while the Y axis corresponds to the proportion (as ratio) of instances belonging to a class value. At

2. http://nsl.cs.unb.ca/NSL-KDD/

Dataset	n. of Instances	n. of Features	n. of Classes
Airlines	539 383	7	2
Electricity	45 312	8	2
Cover Type	581 012	54	7
KDD99	148 517	41	2

Table 1: Dataset characteristics

each moment the sum of the different class labels is equal to 1. We can observe different behaviors as a function of the target variable. Some datasets show high fluctuation of different classes over the stream (*Cover Type* and *Electricity*) while other ones show more smooth (or stable) behavior (*airlines* and *KDD99*).

We think that this final set of four datasets is a good benchmark for evaluating the performance of our approach, DBALSTREAM, w.r.t. state of the art methods.

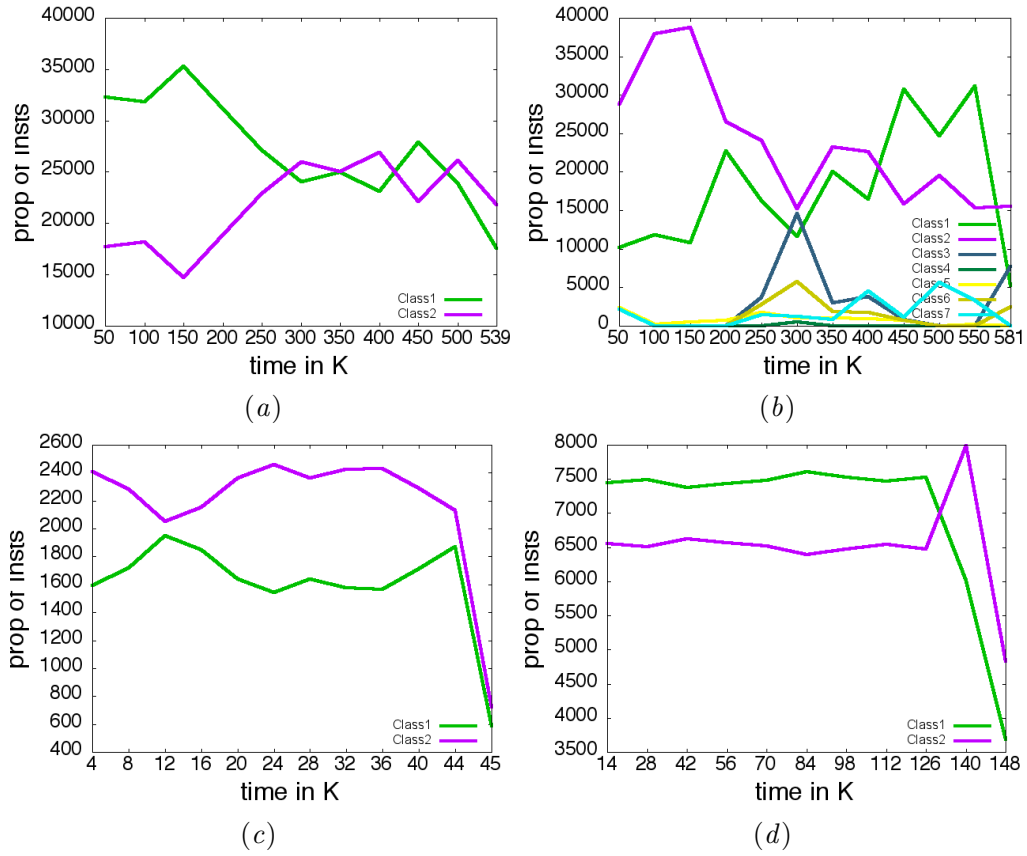


Figure 1: Distribution of classes during the stream process: a) *Airlines* b) *Cover Type* c) *Electricity* e) *KDD99*

4.2. Analysis of Classification Performance

The final accuracy is reported in Figure 2. In this experiment we evaluate the different methods, over the different datasets, varying the budget percentage. We start with a budget percentage of 0.03 and go up to a percentage of 0.3. Obviously, to evaluate the results we need to take into account how the performances change when varying the budget size.

We can observe that DBALStream outperforms the competitors over *Airlines* and *KDD99* datasets while it obtains comparable accuracy results for the other two datasets. In particular, regarding *Cover Type* we can note that all the methods have very similar behavior for budget values smaller or equal to 0.15 while for higher budget the performance of *ACLStream* drops down w.r.t. the competitors. For the *Cover Type* dataset the best result is obtained by DBALStream with a budget value of 0.25. If we analyze the results over *Electricity* we can note that DBALStream has better performance than the other strategies for a big range of the budget value (between 0.1 and 0.25) while for smaller budgets *Rand Unc* slightly outperforms our proposal.

To wrap up this set of experiments we can state that our new proposal obtains better or comparable results with respect to previous approaches in the field of active learning in data streams. These results underline that using the local density information of incoming instances positively affects the active learning process in the data stream setting.

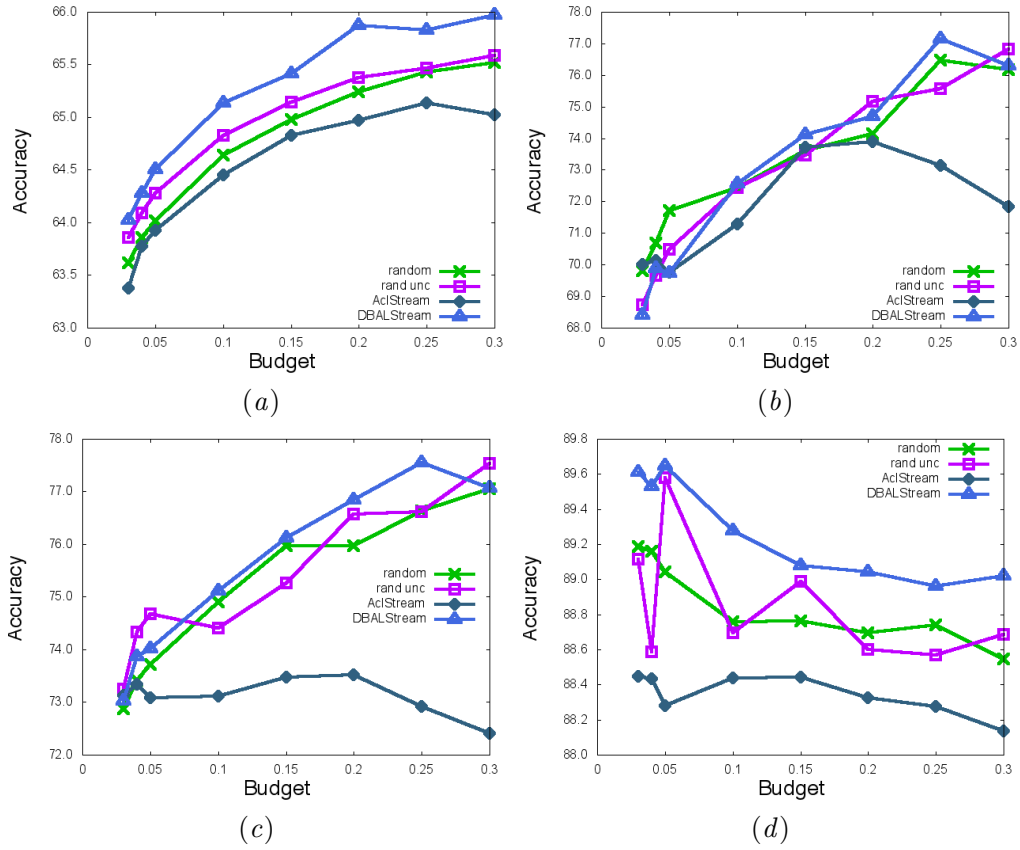


Figure 2: Accuracy on a) Airlines b) Cover Type c) Electricity and e) KDD99

4.3. Influence of the Window Size

In this subsection we focus our attention on the impact of window size over the final performance. In particular, we analyse how the size of the window impacts the final accuracy of DBALSTREAM. For this purpose, we run experiments varying the size of the batches from 50 to 300 with a step of 50. We average each result over 30 runs.

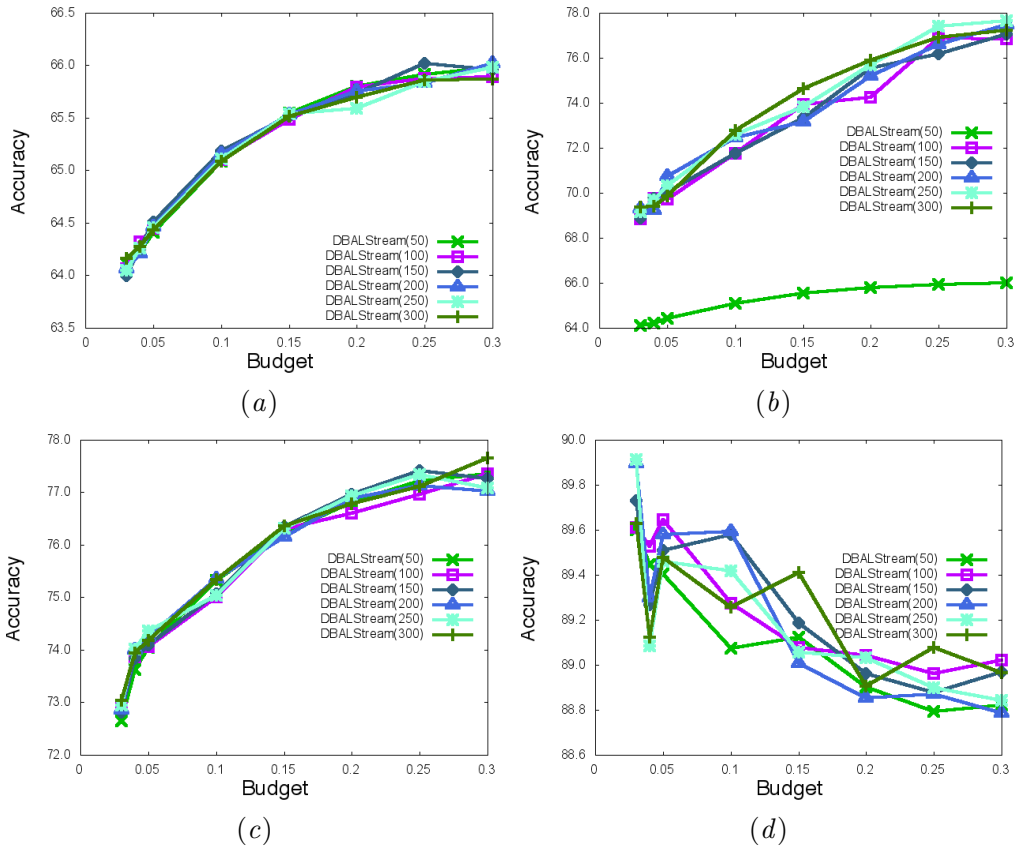


Figure 3: Accuracy Results for DBALSTREAM on a) *Airlines* b) *Cover Type* c) *Electricity* and d) *KDD99* varying the batch size from 50 to 300

The accuracy results are reported in Figure 3. We can observe that, considering *Airlines* and *Electricity* benchmarks the performances of DBALSTREAM are not influenced by this parameter and changes in the value of window size did not impact the final results. A different behavior is shown for the *Cover Type* dataset. In this case we can note an important difference in the results between the smallest value of window size (50) w.r.t. the other values of window size (100-300). This is due to the fact that only considering the 50 previous examples to evaluate the importance of a new one (considering its relative density) is not enough. This can also be explained by the way the data were collected. As we will underline in Section 4.5, the data came from a GIS database that stores raster cells considering their spatial order. These cells are arranged in a matrix and, probably, a window size of 50 did

not allow to capture temporal correlations among the data. For the *KDD99* dataset we note that no particular value of window size outperforms the others. Accuracy fluctuation, also in this case, are very small and they did not exceed 0.5 accuracy points.

In summary, we can underline that for almost all of the benchmarks, this parameter did not affect the final results and window size values bigger than 100 are a good choice in order to compute good density estimates. As a conclusion we can state that, generally, our method is quiet robust with respect to its parameter setting.

4.4. DBALStream: *Confidence* based vs *Margin* based uncertainty estimation

In Section 3 we underlined that our proposal quantifies the uncertainty over an instance considering the *Margin* of the classifier while previous techniques for active learning in data stream exploit the *Confidence* value (Zliobaite et al., 2014). In this section we evaluate how the performances of DBALStream are influenced considering these two different ways of estimating the uncertainty of a classifier over an instance. Figure 4 shows the results of this experiment over all benchmarks datasets.

We can note that, generally, the *Margin* based version clearly outperforms the *Confidence* based one for almost all the datasets and all the budget values. This phenomena is usually more visible for high budget values.

As a general finding we can state that considering the discrepancy between the two most probable classes supplies much more information on the uncertainty of an instance than only taking into account the maximum a posteriori probability. This also supports our choice of using a *Margin*-based approach in our framework.

4.5. Performance with different types of concept drift

For analysis purposes, we also introduce two more datasets: *Cover Type Sorted*, in which the instances of the *Cover Type* dataset are reordered w.r.t. the attribute *elevation*, and *Cover Type Shuffled*, where the order of instances is randomly shuffled. Due to the nature of the underlying problem, sorting the instances by the *elevation* attribute induces a natural gradual drift on the class distribution, because at higher elevation some types of vegetation disappear while other types of vegetation appear gracefully. Shuffling instances ensures that the data distribution is uniform over time, i.e. there is no drift at all. Figure 5 plots the prior probabilities of the classes over time in the two new versions of the dataset

Note, that the order of the *Cover Type* dataset is related to the spatial location of the instances, as clarified by the authors in personal communication. This additional knowledge about the data indicates that in the first two cases (*Cover Type* and *Cover Type Sorted*) the data contains some kind of spatial auto-correlation while in the last scenario (*Cover Type Shuffled*) this autocorrelation is broken by randomisation over the instances order.

Figure 6 plots the accuracy results of the compared approaches on three versions of the Cover Type dataset: the original, sorted (gradual drifts), and shuffled (no drift).

We can observe that all the algorithms obtain similar performances on the original data (Figure 6(a)) while on the sorted version (Figure 6(b)) we can note that DBALStream obtains a slight improvement w.r.t. the competitors for budget values around 0.1.

Considering the *Cover Type Shuffled* version (Figure 6(c)), DBALStream clearly outperforms all the competitors for all the budget values. The only exception happens for

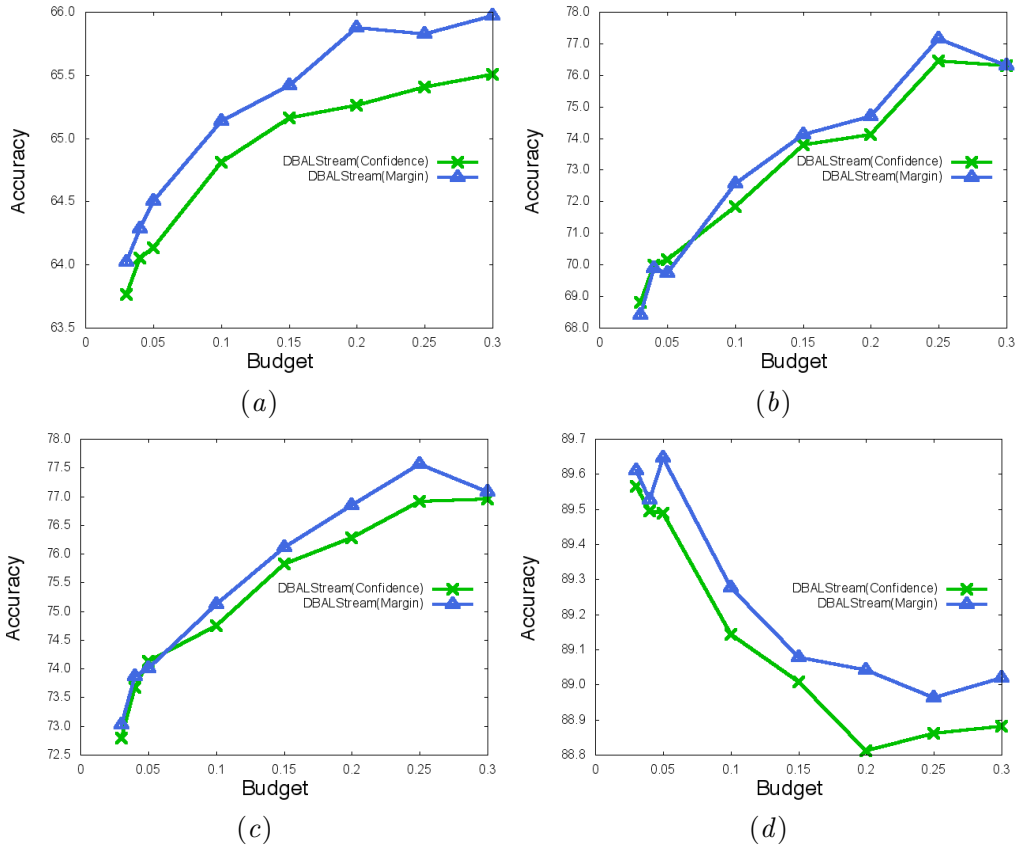


Figure 4: Accuracy Results for DBALSTREAM coupled with different *Confidence* based and *Margin* based uncertainty estimation on a) *Airlines* b) *Cover Type* c) *Electricity* and d) *KDD99*

budget value equals to 0.05 where the *Rand Unc* heuristic obtains the same results as our proposal. This experiment underlines that DBALSTREAM performs similarly to state of the art approaches for spatially auto-correlated data while it clearly outperforms the competitors when data instances are not affected by this kind of auto-correlation.

5. Conclusions

Building classification models on data streams considering only a limited amount of labeled data is becoming a common task due to time and cost constraints.

In this paper we presented DBALSTREAM, a novel algorithm to perform active learning in a data stream scenario. Our approach exploits local instance correlation over the feature space in order to improve sampling on the potentially most useful examples to label.

We assessed the performance of our proposal over real world benchmark datasets. The results showed that the proposed approach outperforms state-of-the-art active learning strategies for data streams in terms of predictive accuracy. We empirically studied dif-

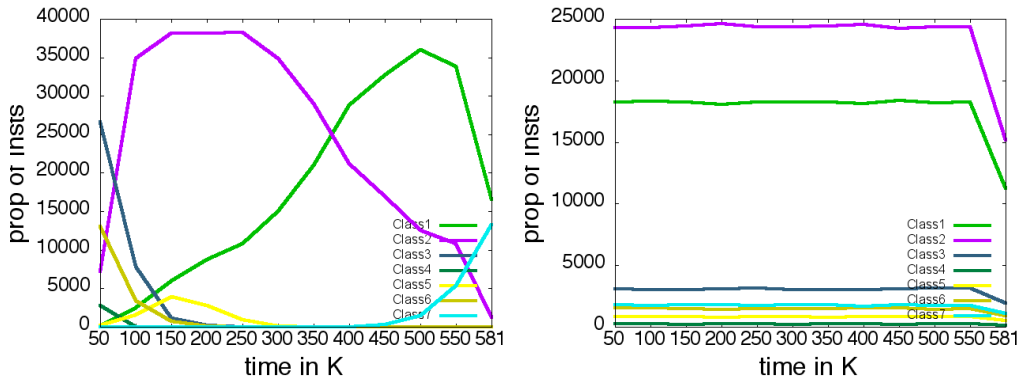


Figure 5: Distribution of the class values during the stream process: a) *Cover Type Sorted* b) *Cover Type Shuffled*

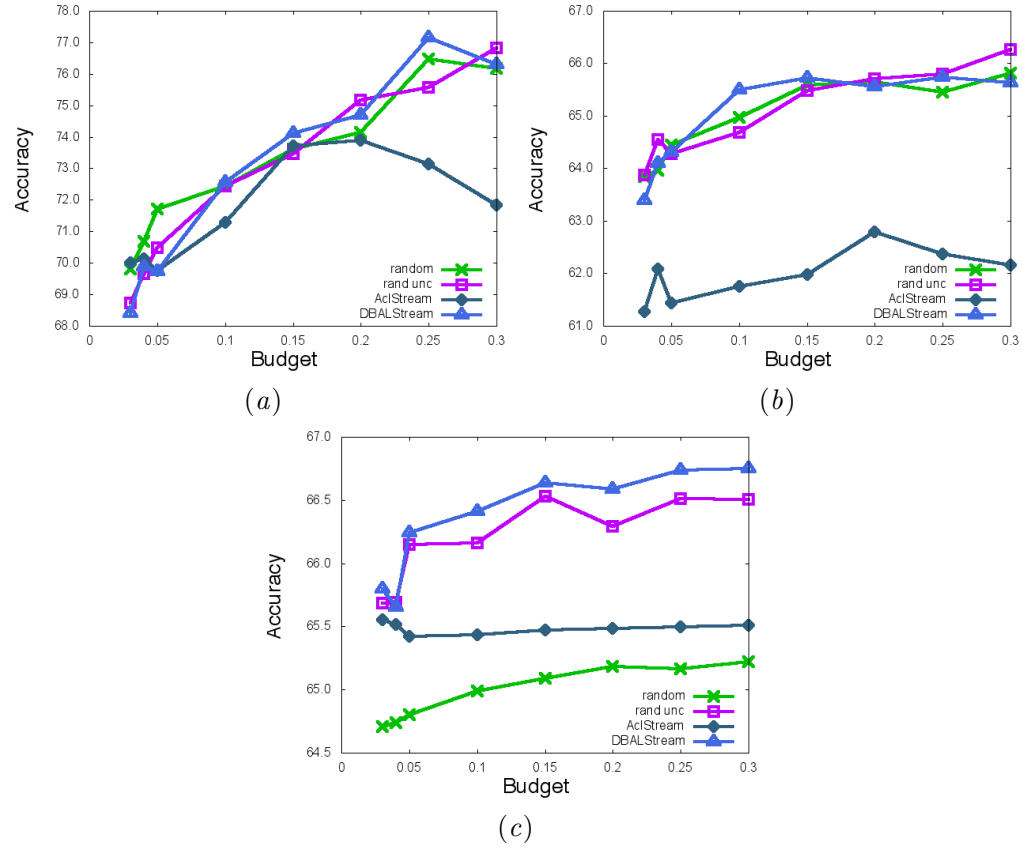


Figure 6: Accuracy on a) *Cover Type* b) *Cover Type Sorted* and c) *Cover Type Shuffled*

ferent factors that can impact our strategy: the window size used to estimate the local density of a new instance and the way in which the uncertainty of an example is estimated. As a final experiment we supplied an in-depth study on how our strategies deal with evolving data and manage concept drift.

As future work we would like to investigate other ways to measure and incorporate density into the active learning process for data streams.

References

J. Attenberg and F. Provost. Online active inference and learning. In *Proc. of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, KDD, pages 186–194, 2011.

K. Bache and M. Lichman. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2013.

A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis. *J. Mach. Learn. Res.*, 11(May):1601–1604, 2010.

N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Worst-case analysis of selective sampling for linear classification. *J. Mach. Learn. Res.*, 7:1205–1230, 2006.

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 1994.

W. Fan, Y. Huang, H. Wang, and Ph. Yu. Active mining of data streams. In *Proc. of SIAM Int. Conf. on Data Mining*, SDM, pages 457–461, 2004.

Y. Fu, X. Zhu, and B. Li. A survey on instance selection for active learning. *Knowl. Inf. Syst.*, 35(2):249–283, 2013.

J. Gama, P. Medas, G. Castillo, and P. Rodrigues. Learning with drift detection. In *Proc. of the 17th Brazilian Symp. on Artificial Intelligence*, SBIA, pages 286–295, 2004.

M. Harries, C. Sammut, and K. Horn. Extracting hidden context. *Machine Learning*, 32 (2):101–126, 1998.

D. Helmbold and S. Panizza. Some label efficient learning results. In *Proc. of the 10th Annual Conf. on Computational Learning Theory*, COLT, pages 218–230, 1997.

Sh. Huang and Y. Dong. An active learning system for mining time-changing data streams. *Intelligent Data Analysis*, 11:401–419, 2007.

D. Ienco, A. Bifet, I. Zliobaite, and B. Pfahringer. Clustering based active learning for evolving data streams. In *Proc. of the 16th Int. Conf. on Discovery Science*, DS, pages 79–93, 2013.

E. Ikonomovska, J. Gama, and S. Dzeroski. Learning model trees from evolving data streams. *Data Mining and Knowledge Discovery*, 23(1):128–168, 2010.

R. Klinkenberg. Using labeled and unlabeled data to learn drifting concepts. In *Proc. of IJCAI workshop on Learning from Temporal and Spatial Data*, pages 16–24, 2001.

D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Proc. of the 17th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR, pages 3–12, 1994.

P. Lindstrom, S. J. Delany, and B. MacNamee. Handling concept drift in a text data stream constrained by high labelling cost. In *Proc. of the 23rd Florida Artificial Intelligence Research Society Conference*, FLAIRS, 2010.

M. Masud, J. Gao, L. Khan, J. Han, and B. Thuraisingham. Classification and novel class detection in data streams with active mining. In *Proc. of the 14th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining*, PAKDD, pages 311–324, 2010.

M. Masud, C. Woolam, J. Gao, L. Khan, J. Han, K. Hamlen, and N. Oza. Facing the reality of data stream classification: coping with scarcity of labeled data. *Knowl. Inf. Syst.*, 33 (1):213–244, 2011.

A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 344 (6191):1492–1496, 2014.

D. Sculley. Online active learning methods for fast label-efficient spam filtering. In *Proc. of the 4th Conf. on Email and Anti-Spam*, CEAS, 2007.

B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison, 2010.

N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic. The role of hubness in clustering high-dimensional data. *IEEE Trans. Knowl. Data Eng.*, 26(3):739–751, 2014.

D. Widyantoro and J. Yen. Relevant data expansion for learning concept drift from sparsely labeled data. *IEEE Trans. on Know. and Data Eng.*, 17:401–412, 2005.

P. Zhao and S. C. H. Hoi. Cost-sensitive online active learning with application to malicious URL detection. In *Proc. of the 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, KDD, pages 919–927, 2013.

X. Zhu, P. Zhang, X. Lin, and Y. Shi. Active learning from data streams. In *Proc. of the 7th IEEE Int. Conf. on Data Mining*, ICDM, pages 757–762, 2007.

I. Zliobaite, A. Bifet, B. Pfahringer, and G. Holmes. Active learning with drifting streaming data. *IEEE Trans. on Neural Networks and Learning Systems*, 25:27–39, 2014.

Appendix C

Combining Active and Transductive Learning

Combining Transductive and Active Learning to Improve Object-based Remote Sensing Image Classification

Fabio Guttler[†], Dino Ienco[†], Pascal Poncelet[*], Maguelonne Teisseire[†]
[†]UMR TETIS, IRSTEA, Montpellier, France
{fabio.guttler, dino.ienco, maguelonne.teisseire}@teledetection.fr
[*]LIRMM (CNRS - Univ. de Montpellier), Montpellier, France
pascal.poncelet@lirmm.fr

***Abstract*—In this letter we propose a new Active Transductive Learning (ATL) framework for object-based classification of remote sensing images. The proposed framework couples a graph based label propagation method with active learning to exploit positive aspects of both learning settings. The transductive approach considers both labeled and unlabeled image objects to perform its classification as they are all available at training time while the active learning strategy smartly guides the construction of the training set employed by the learner. We experiment the proposed framework on two remote sensing datasets coming from the same study area. We compare our proposals w.r.t. state of the art classification techniques employed in object-based image classification.**

I. Introduction

Data labels are usually difficult and expensive to obtain. Standard classification techniques heavily rely on the hypothesis that a big quantity of labeled examples (training set) is available in order to build predictive models. Considering the remote sensing domain, in particular the object-based image classification, the label acquisition task is an important issue because the expert needs to spend time and effort for labeling a portion of the objects to successively classify the rest of the image. The collection of such labels can affect negatively the image classification task on two points of view: the quantity of labeled data commonly needed by standard inductive classifier and the way the objects to label are chosen.

Classical supervised classification approaches (i.e. SVM, Naive Bayes, Random Forest, etc.) require many labeled data to train the model. Also, they assume that training and test data are not available at the same time since the model they have learnt needs to be enough general to classify new unseen examples available in a near future [18]. Conversely, in the case of remote sensing image classification, the available training examples are limited to a small portion of the image and all the objects (training and test) are available at the same time.

A different classification setting is supplied by transductive learning [9]. Transductive learning belongs to the family of semi-supervised approaches and it tries to propagate information from the labeled data to unlabeled one leveraging the availability of training and test data at the same time. These kind of techniques offer an effective approach to supply contextual classification of unlabeled examples by using a relatively small set of labeled ones. Many real-world applications can be modeled through a transductive setting and, in particular, it has been also applied in the remote sensing domain [2] where, labels are difficult to obtain and the classification decisions should not be made separately from learning the current data. Differently from inductive classification, transductive learning does not produce any reusable model, the classification cannot be performed on new data. This is not necessary a problem for object-based image classification as the model learnt on a given image is rarely reused to classify another image [17].

The second issue regards the way labels are collected. Objects are usually labeled randomly from an expert while choosing examples guided by the classifier needs can drastically improve the classification performance and positively impact the data labeling process [5]. This kind of technique is called Active Learning and it allows to involve expert interaction during the classifier construction. More in detail, the choice of the objects to label is guided by the learner needs and the classifier asks the true labels for those objects that can, potentially, improve the performance of the model. The objects are usually selected considering the classifier uncertainty over the set of possible examples to choose [6]. In remote sensing applications this approach is getting more and more attention [5] due to the improvement it can supply in the task of multi and hyperspectral image classification [5], [16].

In this letter we propose to couple transductive and active learning in order to design a new *A*ctive *T*ransductive *L*earning (ATL) framework. More in detail we propose to adapt a label propagation approach [14] to object-based image classification and combine it with an effective active learning strategy. The proposed methodology is experimented in the context of object-based image classification as it is particularly indicated to process high and very high spatial resolution images [1]. In a general way, it starts with a segmentation step which creates a new and more meaningful representation of the image. Instead of an arbitrary pixel grid, segmentation aims to create spatially coherent objects based on spectral and spatial features of adjacent pixels over the image.

The letter is organized as follows: Section II presents the Active Transductive Learning (ATL) framework introducing the Transductive Learner (Sec.II-B) and the Active Learning strategy we adopt (Sec.II-C). The experiments are carried out in Section III: we describe the study area and the datasets (Sec.III-A), the experimental setting (Sec. III-B) and we discuss the obtained results (Sec. III-C). Conclusions are drawn in Section IV.

II. Methodology

In this section we introduce the different components we have used to implement the Active Transductive Learning (ATL) framework : i) the transductive setting ii) the label propagation algorithm we adapted for object-based image classification iii) how we coupled transduction and active learning to perform the final classification.

A. Transductive setting

Given a set of object $O = \{o_i\}_{i=1}^{N}$, let us denote with L the subset of labeled objects of O, and with $U = O \setminus L$ the subset of unlabeled ones. Note that U can in principle have any proportion w.r.t. L, but in many real cases U is much larger than L. Every object in L is assigned a label that refers to one of the known M classes $\mathcal{C} = \{C_j\}_{j=1}^{M}$. We also denote with $\mathbf{Y}$ a $N \times M$ matrix such that $\mathbf{Y}_{ij} = 1$ if C_j is the label assigned to object o_i, 0 otherwise. Without loss of generality, we can refer to L as training data and to U as test data.

The goal of a *transductive learner* is to make an inference "from particular to particular", i.e. given the classifications of the instances in the training set L, it aims to predict the classifications of the instances in the test set U, rather than inducing a general rule that works out for classifying new unseen instances [18]. Transduction is naturally related to the class of case-based learning algorithms, whose most well-known algorithm is the k-nearest neighbor (kNN) [10]. Differently from standard supervised setting, in the transductive setting there is no separation between model training and testing phase. The classification of new unseen example is performed at the same time the model is learnt over L.

B. Label propagation algorithm

In order to perform transductive learning we use the approach proposed in [14] named *R*obust *Multi*-class *Graph T*ransduction (*RMGT*). From the best of our knowledge it is the first time this approach is employed in a remote sensing application and in particular to perform an object-based classification of satellite images.

Essentially, *RMGT* implements a graph-based label propagation approach, which exploits a kNN graph built over the entire dataset to propagate the class information from the labeled to the unlabeled examples.

The assumption behind this approach is that adjacent vertices are likely to have similar labels. For this reason the label propagation procedure ensures that the classification function varies smoothly along the edges of the kNN graph. In the following we describe in detail the mathematics aspects of *RMGT*.

Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, w \rangle$ be an undirected graph whose vertex set is $\mathcal{V} = O$, edge set is $\mathcal{E} = \{(o_i, o_j) | o_i, o_j \in O \wedge sim(o_i, o_j) > 0\}$, and edge weighting function is $w = sim(o_i, o_j)$.

Given a positive integer k, consider the kNN graph $\mathcal{G}_k = \langle \mathcal{V}, \mathcal{E}_k, w \rangle$ derived from $\mathcal{G}$ and such that $\mathcal{E} = \{(d_i, d_j) | d_j \in N_i\}$, where N_i denotes the set of d_i's k-nearest neighbors. A weighted sparse matrix is obtained as $\mathbf{W} = \mathbf{A} + \mathbf{A}^{\mathrm{T}}$, where $\mathbf{A}$ is the weighted adjacency matrix of $\mathcal{G}_k$ and $\mathbf{A}^{\mathrm{T}}$ is the transpose of $\mathbf{A}$; the matrix $\mathbf{W}$ represents a *symmetry-favored* kNN graph [14]. Moreover, let $\mathbf{L} = \mathbf{I}_N - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ the normalized Laplacian of $\mathbf{W}$, where $\mathbf{I}_N$ is the $N \times N$ identity matrix and $\mathbf{D} = diag(\mathbf{W}\mathbf{1}_N)$. Without loss of generality, we can rewrite $\mathbf{L}$ and $\mathbf{W}$ as subdivided into four and two submatrices, respectively:

$$\mathbf{L} = \begin{bmatrix} \mathbf{\Delta}_{\mathcal{L}\mathcal{L}} & \mathbf{\Delta}_{\mathcal{L}\mathcal{U}} \\ \mathbf{\Delta}_{\mathcal{U}\mathcal{L}} & \mathbf{\Delta}_{\mathcal{U}\mathcal{U}} \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} \mathbf{Y}_{\mathcal{L}} \\ \mathbf{Y}_{\mathcal{U}} \end{bmatrix} \tag{1}$$

where $\mathbf{\Delta}_{\mathcal{L}\mathcal{L}}$ and $\mathbf{Y}_{\mathcal{L}}$ are the submatrices of $\mathbf{L}$ and $\mathbf{Y}$, respectively, corresponding to the labeled objects, and analogously for the other submatrices. The RMGT learning algorithm finally yields a matrix $\mathbf{F} \in \mathbb{R}^{N \times M}$ defined as:

$$\mathbf{F} = -\mathbf{\Delta}_{\mathcal{U}\mathcal{U}}^{-1}\mathbf{\Delta}_{\mathcal{U}\mathcal{L}}\mathbf{Y}_{\mathcal{L}} + \frac{\mathbf{\Delta}_{\mathcal{U}\mathcal{U}}^{-1}\mathbf{1}_u}{\mathbf{1}_u^{\mathrm{T}}\mathbf{\Delta}_{\mathcal{U}\mathcal{U}}^{-1}\mathbf{1}_u}(N\omega - \mathbf{1}_l^{\mathrm{T}}\mathbf{Y}_{\mathcal{L}} + \mathbf{1}_u^{\mathrm{T}}\mathbf{\Delta}_{\mathcal{U}\mathcal{U}}^{-1}\mathbf{\Delta}_{\mathcal{U}\mathcal{L}}\mathbf{Y}_{\mathcal{L}}) \tag{2}$$

where $\omega \in \mathbb{R}^M$ is the class prior probabilities.

The transductive learning scheme used by RMGT employs spectral properties of the kNN graph to spread the labeled information over the set of test instances. Specifically, the label propagation process is modeled as a constrained convex optimization problem where the labeled objects are employed to constrain and guide the final classification. Equation 2 represents the closed form solution of the propagation process. This equation shows how Labeled ($\mathcal{L}$) and Unlabeled ($\mathcal{U}$) examples are combined in order to implement the main assumption that adjacent vertices are likely to have similar labels. After the propagation step, every unlabeled object o_i is associated to a vector (i.e., the i-th row of $\mathbf{F}$) representing the likelihood of the object o_i for each of the classes; therefore, o_i is assigned to the class that maximizes the likelihood. Concerning the $sim(\cdot, \cdot)$ function, we derived it from the standard euclidean distance. In particular the $sim(\cdot, \cdot)$ is defined as $\frac{1}{1+dist(\cdot,\cdot)}$ where $dist(\cdot, \cdot)$ is the euclidean distance between the feature vectors of two objects. Moreover, the class priors (ω) used in Eq. (2) are defined as uniformly distributed.

Algorithm 1 sketches the main steps of the Label Propagation algorithm. Initially, the similarity matrix between all the objects is computed (Line 1). The computation of the similarity matrix is based on the euclidean distance measure. The graph-based label propagation process requires the construction of the kNN graph (Line 2) and its symmetry-favored transformation (Line 3). After that, the algorithm computes the normalized Laplacian of the matrix (Line 4) and the $RMGT$ algorithm is applied on such data matrix. Line 6 describes the decision rule we adopted to perform the classification.

Algorithm 1 Object-Based Transductive Classification

Input: A collection of object O, with labeled objects L and unlabeled objects U (with $D = L \cup U$ and $L \cap U = \emptyset$); a set of labels $\mathcal{C} = \{C_j\}_{j=1}^{M}$ assigned to the objects in L; a positive integer k for the neighborhood selection.
Output: A classification over $\mathcal{C}$ for the objects in U.
1: Build the similarity graph $\mathcal{G}$ for the object set O.
2: Extract the k-nearest neighbor graph $\mathcal{G}_k$ from $\mathcal{G}$. */* Section II-B */*
3: Build the matrix $\mathbf{W}$ from $\mathcal{G}_k$, which represents the symmetry-favored k-nearest neighbor graph. */* Section II-B */*
4: Compute the normalized Laplacian of $\mathbf{W}$. */* Section II-B */*
5: Compute the *RMGT* solution $\mathbf{F}$. */* Eq. (2) */*
6: Assign object $o_i \in U$ to the class C_{j*} that maximizes the class likelihood, $j^* = \arg\max_j \mathbf{F}_{ij}$.

C. Active Learning

Active learning is getting more attention in the remote sensing image classification domain as it helps to deal with the time and effort consuming task of collecting a good quality training set to build a classification model [4], [5]. The general active learning loop [6] involves the interaction between the classifier and the expert. Firstly a budget is defined. It represents the percentage (or the number) of examples the experts is willing to label. Then the active learning loop starts. At each iteration the active learning procedure ranks the set of unlabeled examples in order to promote in the rank the more relevant ones to label. The rank is produced scoring each example with its importance considering the current learnt classifier. Once the rank is produced, the procedure chooses the top objects (one or more) and provide them to the expert in order to obtain the true labels. The new objects are added to the current training set and the classifier is updated. The active learning cycle stops when the budget is exhausted.

Different heuristics to score the examples were proposed in the literature but in this work we chose the *Margin-based* strategy [6]. This strategy considers the probability distribution of a classifier cl on the example x over the possible set of classes $\mathcal{C}$ and it is prone to select instances with minimum margin between posterior probabilities of the two most likely class labels. More formally, it is defined as follows:

$$Margin(x) = P_{cl}(x|C_i) - P_{cl}(x|C_j)$$

where C_i is the most probable class for the example x while C_j is the second most probable class given the classifier cl.

Values of $Margin(x)$ close to 0 indicates big uncertain on x while values close to 1 underlines reliable confidence in the prediction. In the active learning step, first the unlabeled instances are ranked in ascending order w.r.t. their *Margin* value, then the top n examples are supplied to the expert and their true label is obtained. In our application we fixed the number of examples at each loop equals to 20.

Considering our framework, we coupled the *Margin-based* heuristic with Algorithm 1. Given an object to classify o_i, we employ the likelihood vector returned by $RMGT$ ($\mathbf{F}_i$) as posterior probability distribution to implement the *Margin-based* strategy.

III. Experiments

A. Dataset Description

Experiments were performed on the *Lower Aude Valley* site, located in the south of France. Spanning over a coastal wetland area of about 4 842 ha, this site is part of the European network of protection areas called Natura 2000. Most of the site (56.3 %) is composed of natural and semi-natural areas, specially salt-meadows, salt-marshes and coastal lagoons. The rest of the site (43.7 %) is principally occupied by agricultural parcels (vineyards, cereal crops, orchards) and some small artificial areas (roads, houses).

As input remote sensing data we have used a RapidEye multispectral image (6.5 m of spatial resolution) acquired in 24 June 2009 and available in the context of the GEOSUD project (France). The RapidEye image contains five spectral bands (approximate center in nm): blue (475), green (555), red (657), red-edge (710) and near-infrared (805). Image segmentation was performed using the five spectral bands and considering only the area inside the Lower Aude Valley site. The segmentation task has created a set of 13,292 objects (using the multiresolution segmentation algorithm MSA available at *eCognition Developer 8.8.1*). For each object we have calculated the following attributes: mean value for the 5 spectral bands and 5 spectral indices ($NDVI$ [15], $NDVI_{re}$ [7], [13], $NDWI_{re}$ [12], $RTVI_{core}$ [3] and SR [11]).

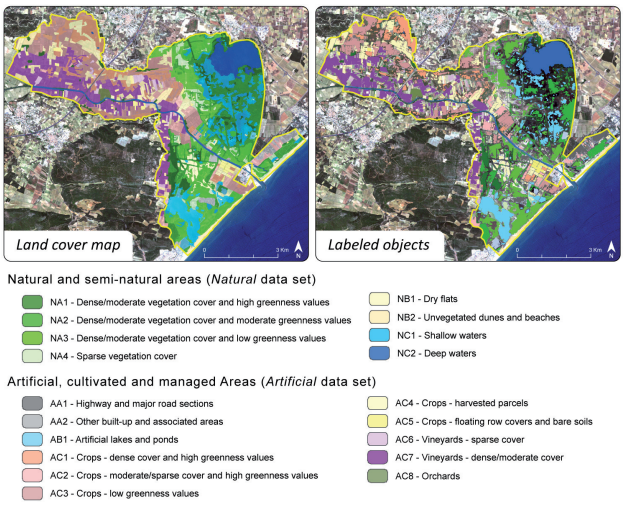


Fig. 1. Land cover map for the *Lower Aude Valley* site (left side) and the spatial distribution of the labeled objects (right side).

In parallel, the same RapidEye image was used to map the whole *Lower Aude Valley* site. This task was carried out through a manual land cover digitalization process at the scale of 1:10,000. Field surveys and precise aerials photographs (0.5 m of spatial resolution) were employed to ensure the exactness of the land cover map. Each individual map unit (polygons in our case) was labeled according to two specific sets of land cover classes. The first set is specific to natural and semi-natural areas (*Natural* set) while the second set concerns artificial, cultivated and managed areas (*Artificial* set). Then,

the land cover map was superimposed on the set of objects in order to propagate the land cover information. Only the objects fitting completely inside the polygons of the map received the land cover label (see figure 1). In total, 3 357 objects were labeled for the *Natural* set and 3 637 for the *Artificial* set.

As the experiment reproduces a real task of land cover object-based classification, the number of objects per class is strongly unbalanced as one can notice in the following lists (number of objects indicated in brackets).

Natural: NA1(264), NA2(760), NA3(1019), NA4(161), NB1(253), NB2(155), NC1(529) and NC2(216).

Artificial: AA1(21), AA2(418), AB1(77), AC1(439), AC2(277), AC3(658), AC4(542), AC5(209), AC6(75), AC7(900) and AC8(21).

B. Experimental Setting

We compared our proposal with respect to state of the art classification approaches. As competitors we used the *Random Forest* Classifier (RF), the *Support Vector Machine* (SVM) and the *Naive Bayes* approach (NB). For *SVM* we used Polynomial kernel with exponent value equals to 8. We coupled each of the previous classifiers with the same active learning strategy we employed for ATL. This is done in order to fairly compare our proposal with the competitors. We also investigated the performance of the base Transductive Learner (TL) w.r.t. ATL to highlight the benefit supplied by the Active Learning step.

For all the competitors we use the *Weka*[1] implementation with default setting. For the $RMGT$ method we use a k value equals to 15 for building the kNN graph.

We varied the training percentage (budget) between 2% to 40% in steps of 2%. The percentage indicates the proportion of the original dataset employed as training set.

We evaluated the classification performance using the *F-Measure* [8]. We used *F-Measure* instead of general accuracy due to its ability to better describe classifier performance on unbalanced dataset. We randomly initialised each classifier with an object per class and then the active learning process starts. For each pair 'classifier and training percentage' we reported the average results over 30 runs.

C. Experimental Results

Figure 2 and Figure 3 report the results over the *Natural* and *Artificial* subsamples of the *Lower Aude Valley* site.

Considering the results on the *Natural* subsample reported in Figure 2, we can observe that the general trend is that by increasing the number of objects available in the training set the performance increases. Comparing the different classification methods we can observe that ATL outperforms all the other competitors for all the values of training percentage. The biggest gap between ATL and the other methods is obtained for a training percentage of 4% where it obtains more than 4 points of *F-Measure* w.r.t. the *Random Forest* method. For all the other values of training percentage the gain is always around 1.5 or 3 points of *F-Measure*. The maximum value of *F-Measure* for our proposal (0.78) is reached for a training percentage of 40%.

Figure 2(b) reports the performance of the Transductive approach with and without Active Learning. We can observe that for training percentages bigger than 8% ATL clearly outperforms the simple Transductive Learner underlining that building a training set guided by the classifier needs positively influences the final performance. Regarding smaller training percentages, we can note that the difference is very limited. This fact points out that the benefit of Active Learning only becomes evident when the size of training set exceeds a minimum threshould, in our case around 8%.

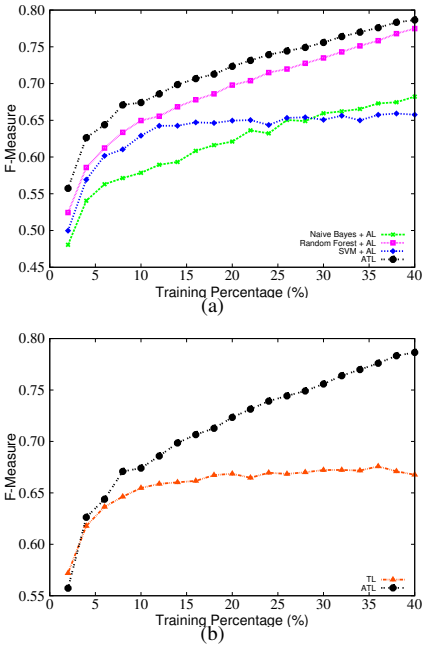


Fig. 2. Results of F-Measure a) All the different classification methods coupled with Active Learning b) ATL vs TL over the *Natural* subsample of *Lower Aude Valley* site.

Figure 3 reports the results of the different approaches over the *Artificial* subsample of the *Lower Aude Valley* study area.

We can observe that, for percentage training smaller than 20%, $SVM + AL$ obtains results that differ of 2 or 3 points from ones reached by ATL . When the budget increases and reaches values bigger than 20% the general trend changes, ATL continues to improve its performance outperforming *SVM* that remains stable. The maximum gap between ATL and *SVM* is around 7 points achieved for a budget of 40%.

Comparing the performance of ATL w.r.t. *Naive Bayes* and *Random Forest* we can note that, also in this case, ATL outperforms both approaches. This experiment shows that for reasonable amount of training percentage available (20%) the ATL framework is able to clearly outperforms all the state of the art methods.

Figure 3(b) shows the comparison between ATL and TL. The behaviour of the two approaches follow the same trend showed for the *Natural Lower Aude Valley* dataset. ATL

[1] http://www.cs.waikato.ac.nz/ml/weka/

clearly outperforms TL for training percentage bigger than 8% while for lesser number of labeled instances ATL and TL obtain comparable performance.

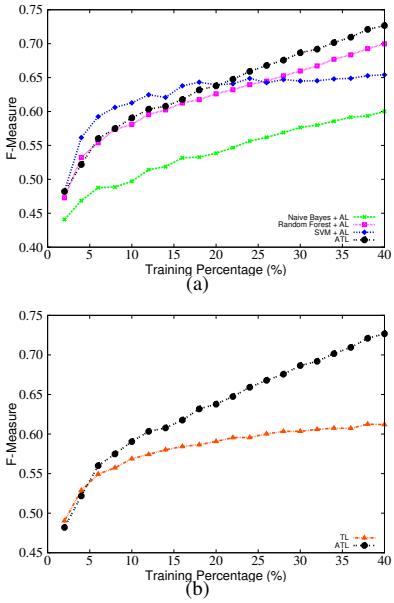


Fig. 3. Results of F-Measure a) All the different classification methods coupled with Active Learning b) ATL vs TL over the *Artificial* subsample of *Lower Aude Valley* site.

To sum up the experimental section, the proposed approach outperforms the competitors over both datasets considering a reasonable amount of training data. It also shows a stable behaviour w.r.t. all the other state of the art classification methods usually employed for object-based image classification over the considered remote sensing datasets. This result can be explained considering the nature of the classifier. Standard classification techniques employ inductive learning where training and test examples are available at separate time. Conversely, transductive learning assumes that training and test data are available at the same time and they can be exploited together to propagate class assignment from labeled to unlabeled data. The unlabeled examples in the training time allow the transductive learner to make an inference "from particular to particular". In the case of remote sensing datasets, where training and test data are available at the same time, this setting can be more adapt. The Active Transductive Learning approach we have proposed also take advantage from the Active Learning strategy to intelligently build the training set.

IV. Conclusion

In this letter we presented a new Active Transductive Learning Framework that can efficiently deal with object-based image classification. The proposed approach was experimented on two remote sensing datasets coming from the same study area (*Lower Aude Valley*). We compared our proposals w.r.t. classification approaches usually employed in object-based image classification. The quality of the obtained results underlines the appropriateness of combining transduction and active learning for the classification of remote sensing images. As future work we would investigate more sophisticated active learning techniques that employ diversity criteria in the selection process to better exploit the structure of the data.

Acknowledgment

The authors would like to acknowledge the French National Research Agency in the framework of the program "Investissements d'Avenir" (GEOSUD project, ANR-10-EQPX-20).

References

[1] T. Blaschke, G.J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R.Q. Feitosa, F. van der Meer, H. van der Werff, F. van Coillie, and D. Tiede. Geographic object-based image analysis towards a new paradigm. *J. of Photogrammetry and Remote Sensing*, 87(0):180 – 191, 2014.

[2] L. Bruzzone, M. Chi, and M. Marconcini. A novel transductive SVM for semisupervised classification of remote-sensing images. *IEEE T. Geoscience and Remote Sensing*, 44(11-2):3363–3373, 2006.

[3] P.-F. Chen, N. Tremblay, J.-H. Wang, P. Vigneault, W.-J. Huang, and B.-G. Li. New index for crop canopy fresh biomass estimation. *Guang Pu Xue Yu Guang Pu Fen Xi/Spectroscopy and Spectral Analysis*, 30(2):512–517, 2010.

[4] B. Demir and L. Bruzzone. A novel active learning method in relevance feedback for content-based remote sensing image retrieval. *IEEE T. Geoscience and Remote Sensing*, 53(5):2323–2334, 2015.

[5] B. Demir, L. Minello, and L. Bruzzone. An effective strategy to reduce the labeling cost in the definition of training sets by active learning. *IEEE Geosci. Remote Sensing Lett.*, 11(1):79–83, 2014.

[6] Y. Fu, X. Zhu, and B. Li. A survey on instance selection for active learning. *Knowl. Inf. Syst.*, 35(2):249–283, 2013.

[7] A. Gitelson and M. N. Merzlyak. Spectral reflectance changes associated with autumn senescence of aesculus hippocastanum l. and acer platanoides l. leaves. spectral features and relation to chlorophyll estimation. *Journal of Plant Physiology*, 143(3):286 – 292, 1994.

[8] L. Gómez-Chova, J. Muñoz-Marí, V. Laparra, J. Malo-López, and G. Camps-Valls. A review of kernel methods in remote sensing data analysis. In *Optical Remote Sensing*, pages 171–206. Springer, 2011.

[9] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, 1999.

[10] T. Joachims. Transductive learning via spectral graph partitioning. In *ICML*, pages 290–297, 2003.

[11] C. F. Jordan. Derivation of leaf-area index from quality of light on the forest floor. *Ecology*, 50(4):pp. 663–666, 1969.

[12] S. Klemenjak, B. Waske, S. Valero, and J. Chanussot. Unsupervised river detection in rapideye data. In *IGARSS*, pages 6860–6863, 2012.

[13] A. Kross, H. McNairn, D. Lapen, M. Sunohara, and C. Champagne. Assessment of rapideye vegetation indices for estimation of leaf area index and biomass in corn and soybean crops. *International Journal of Applied Earth Observation and Geoinformation*, 34:235–248, 2015.

[14] W. Liu and S.-F. Chang. Robust multi-class transductive learning with graphs. In *CVPR*, pages 381–388, 2009.

[15] JW Rouse Jr, RH Haas, JA Schell, and DW Deering. Monitoring vegetation systems in the great plains with erts. *NASA special publication*, 351:309, 1974.

[16] S. Sun, P. Zhong, H. Xiao, and R. Wang. Active learning with gaussian process classifier for hyperspectral image classification. *IEEE T. Geoscience and Remote Sensing*, 53(4):1746–1760, 2015.

[17] Z. Sun, C. Wang, D. Li, and J. Li. Semisupervised classification for hyperspectral imagery with transductive multiple-kernel learning. *IEEE Geosci. Remote Sensing Lett.*, 11(11):1991–1995, 2014.

[18] V. Vapnik. *Statistical learning theory*. Wiley, 1998.

Bibliography

[1] A. A. Abin and H. Beigy. "Active selection of clustering constraints: a sequential approach". In: *Pattern Recognition* 47.3 (2014), pp. 1443–1458.

[2] C. C. Aggarwal, ed. *Data Streams - Models and Algorithms*. Advances in Database Systems. Springer, 2007. ISBN: 978-0-387-28759-1.

[3] Charu C. Aggarwal and Haixun Wang. *Managing and Mining Graph Data*. 1st. Springer Publishing Company, Incorporated, 2010. ISBN: 1441960449, 9781441960443.

[4] M. Akdere et al. "Learning-based Query Performance Modeling and Prediction". In: *IEEE 28th International Conference on Data Engineering (ICDE 2012), Washington, DC, USA (Arlington, Virginia), 1-5 April, 2012*. 2012, pp. 390–401.

[5] M. Akdere et al. "The Case for Predictive Database Systems: Opportunities and Challenges". In: *CIDR 2011, Fifth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 9-12, 2011, Online Proceedings*. 2011, pp. 167–174.

[6] P. Anchuri et al. "Approximate graph mining with label costs". In: *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*. 2013, pp. 518–526.

[7] P. Arcaini et al. "User-driven geo-temporal density-based exploration of periodic and not periodic events reported in social networks". In: *Information Sciences* Accepted for Publication (2016), pp. 122–143.

[8] Y. Bengio, A. C. Courville, and P. Vincent. "Representation Learning: A Review and New Perspectives". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.8 (2013), pp. 1798–1828.

[9] A. Bifet et al. "CD-MOA: Change Detection Framework for Massive Online Analysis". In: *Advances in Intelligent Data Analysis XII - 12th International Symposium, IDA 2013, London, UK, October 17-19, 2013. Proceedings*. 2013, pp. 92–103.

[10] B. Boden et al. "Mining coherent subgraphs in multi-layer graphs with edge labels". In: *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*. 2012, pp. 1258–1266.

[11] V. Bogorny and S. Shekhar. "Spatial and Spatio-temporal Data Mining". In: *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*. 2010, p. 1217.

[12] K. D. Bollacker, R. P. Cook, and P. Tufts. "Freebase: A Shared Database of Structured General Human Knowledge". In: *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*. 2007, pp. 1962–1963.

[13] F. Bonchi, C. Castillo, and D. Ienco. "Meme ranking to maximize posts virality in microblogging platforms". In: *J. Intell. Inf. Syst.* 40.2 (2013), pp. 211–239.

[14] F. Bonchi et al. "Distance oracles in edge-labeled graphs". In: *Proceedings of the 17th International Conference on Extending Database Technology, EDBT 2014, Athens, Greece, March 24-28, 2014.* 2014, pp. 547–558.

[15] V. Bonnici et al. "A subgraph isomorphism algorithm and its application to biochemical data". In: *BMC Bioinformatics* 14.S-7 (2013), S13.

[16] G. Bordogna and D. Ienco. "Fuzzy Core DBScan Clustering Algorithm". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems - 15th International Conference, IPMU 2014, Montpellier, France, July 15-19, 2014, Proceedings, Part III*. 2014, pp. 100–109.

[17] B.n Bringmann et al. "Learning and Predicting the Evolution of Social Networks". In: *IEEE Intelligent Systems* 25.4 (2010), pp. 26–35.

[18] A. Cakmak and G. Özsoyoglu. "Taxonomy-superimposed graph mining". In: *EDBT 2008, 11th International Conference on Extending Database Technology, Nantes, France, March 25-29, 2008, Proceedings*. 2008, pp. 217–228.

[19] F. Cao and J. Z. Huang. "A Concept-Drifting Detection Algorithm for Categorical Evolving Data". In: *PAKDD (2)*. 2013.

[20] V. Chandola, S. Boriah, and V. Kumar. "A Framework for Exploring Categorical Data". In: *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA*. 2009, pp. 187–198.

[21] F. Coenen, P. H. Leng, and S. Ahmed. "Data Structure for Association Rule Mining: T-Trees and P-Trees". In: *IEEE Trans. Knowl. Data Eng.* 16.6 (2004), pp. 774–778.

[22] B. Demir and L. Bruzzone. "Hashing-Based Scalable Remote Sensing Image Search and Retrieval in Large Archives". In: *IEEE T. Geoscience and Remote Sensing* 54.2 (2016), pp. 892–904.

[23] B. Demir, L. Minello, and L. Bruzzone. "Definition of Effective Training Sets for Supervised Classification of Remote Sensing Images by a Novel Cost-Sensitive Active Learning Method". In: *IEEE T. Geoscience and Remote Sensing* 52.2 (2014), pp. 1272–1284.

[24] M. Elseidy et al. "GRAMI: Frequent Subgraph and Pattern Mining in a Single Large Graph". In: *PVLDB* 7.7 (2014), pp. 517–528.

[25] A. Freitas et al. "Approximate and selective reasoning on knowledge graphs: A distributional semantics approach". In: *Data Knowl. Eng.* 100 (2015), pp. 211–225.

[26] Y. Fu, X. Zhu, and B. Li. "A survey on instance selection for active learning". In: *Knowl. Inf. Syst.* 35.2 (2013), pp. 249–283.

[27] J. Gama et al. "Learning with drift detection". In: *SBIA*. 2004.

[28] S. Gautama et al. "Relevance Criteria for Spatial Information Retrieval Using Error-Tolerant Graph Matching". In: *IEEE Trans. Geoscience and Remote Sensing* 45.4 (2007), pp. 810–817.

[29] A. Gionis, P. Indyk, and R. Motwani. "Similarity Search in High Dimensions via Hashing". In: *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK*. 1999, pp. 518–529.

[30] K. Greff et al. "LSTM: A Search Space Odyssey". In: *CoRR* abs/1503.04069 (2015).

[31] F. Guttler et al. "Combining Transductive and Active Learning to Improve Object-based Classification of Remote Sensing Images". In: *Remote Sensing Letters* Accepted for publication.- (2016), pp. 1–10.

[32] F. Guttler et al. "Exploring high repetitivity remote sensing time series for mapping and monitoring natural habitats - A new approach combining OBIA and k-partite graphs". In: *2014 IEEE Geoscience and Remote Sensing Symposium, IGARSS 2014, Quebec City, QC, Canada, July 13-18, 2014*. 2014, pp. 3930–3933.

[33] F. Guttler et al. "Towards the Use of Sequential Patterns for Detection and Characterization of Natural and Agricultural Areas". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems - 15th International Conference, IPMU 2014, Montpellier, France, July 15-19, 2014, Proceedings, Part I*. 2014, pp. 97–106.

[34] P. N. Hai et al. "Extracting Trajectories through an Efficient and Unifying Spatio-temporal Pattern Mining System". In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II*. 2012, pp. 820–823.

[35] P. Nhat Hai et al. "Mining Fuzzy Moving Object Clusters". In: *Advanced Data Mining and Applications, 8th International Conference, ADMA 2012, Nanjing, China, December 15-18, 2012. Proceedings*. 2012, pp. 100–114.

[36] P. Nhat Hai et al. "Mining Representative Movement Patterns through Compression". In: *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part I*. 2013, pp. 314–326.

[37] P. Nhat Hai et al. "Mining time relaxed gradual moving object clusters". In: *SIGSPATIAL 2012 International Conference on Advances in Geographic Information Systems (formerly known as GIS), SIGSPATIAL'12, Redondo Beach, CA, USA, November 7-9, 2012*. 2012, pp. 478–481.

[38] J. Han, Z. Li, and L. A. Tang. "Mining Moving Object, Trajectory and Traffic Data". In: *Database Systems for Advanced Applications*. Vol. 5982. Lecture Notes in Computer Science. 2010, pp. 485–486. ISBN: 978-3-642-12097-8.

[39] M. Al Hasan and M. J. Zaki. "Output Space Sampling for Graph Patterns". In: *PVLDB* 2.1 (2009), pp. 730–741.

[40] T. Hendrickx et al. “Mining Association Rules in Graphs Based on Frequent Cohesive Itemsets”. In: *Advances in Knowledge Discovery and Data Mining - 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings, Part II*. 2015, pp. 637–648.

[41] X. Huang and L. Zhang. “An SVM Ensemble Approach Combining Spectral, Structural, and Semantic Features for the Classification of High-Resolution Remotely Sensed Imagery”. In: *IEEE T. Geoscience and Remote Sensing* 51.1 (2013), pp. 257–272.

[42] M. Hussaina et al. “Change detection from remotely sensed images: From pixel-based to object-based approaches”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 80 (2013), 91—106.

[43] D. Ienco, R. G. Pensa, and R. Meo. “From Context to Distance: Learning Dissimilarity for Categorical Data Clustering”. In: *TKDD* 6.1 (2012), p. 1.

[44] D. Ienco, R. G. Pensa, and R. Meo. “Parameter-Free Hierarchical Co-clustering by n-Ary Splits”. In: *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part I*. 2009, pp. 580–595.

[45] D. Ienco and R.G. Pensa. “Positive and unlabeled learning in categorical data”. In: *Neurocomputing* Accepted for Publication (2016), pp. 1–24.

[46] D. Ienco, R.G. Pensa, and R. Meo. “A Semi-Supervised Approach to the Detection and Characterization of Outliers in Categorical Data”. In: *Transaction on Neural Network and Learning Systems* Accepted for Publication (2016), pp. 1–13.

[47] D. Ienco, I. Zliobaite, and B. Pfahringer. “High density-focused uncertainty sampling for active learning over evolving stream data”. In: *Proceedings of the 3rd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, BigMine 2014, New York City, USA, August 24, 2014*. 2014, pp. 133–148.

[48] D. Ienco et al. “Change detection in categorical evolving data streams”. In: *Symposium on Applied Computing, SAC 2014, Gyeongju, Republic of Korea - March 24 - 28, 2014*. 2014, pp. 792–797.

[49] D. Ienco et al. “Clustering Based Active Learning for Evolving Data Streams”. In: *Discovery Science - 16th International Conference, DS 2013, Singapore, October 6-9, 2013. Proceedings*. 2013, pp. 79–93.

[50] D. Ienco et al. “Parameter-less co-clustering for star-structured heterogeneous data”. In: *Data Min. Knowl. Discov.* 26.2 (2013), pp. 217–254.

[51] H. Jeung et al. “Discovery of convoys in trajectory databases”. In: *Proc. VLDB Endow.* 1.1 (2008), pp. 1068–1080.

[52] R. Jin et al. “Computing label-constraint reachability in graph databases”. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, June 6-10, 2010*. 2010, pp. 123–134.

[53] M. Kuramochi and G. Karypis. "Finding Frequent Patterns in a Large Sparse Graph*". In: *Data Min. Knowl. Discov.* 11.3 (2005), pp. 243–271.

[54] Y. LeCun, Y. Bengio, and G. Hinton. "Deep Learning". In: *Nature* 52.8 (2015), pp. 436–444.

[55] J. Lehmann et al. "DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia". In: *Semantic Web* 6.2 (2015), pp. 167–195.

[56] W. Li, Q. Guo, and C. Elkan. "A Positive and Unlabeled Learning Algorithm for One-Class Classification of Remote-Sensing Data". In: *IEEE T. Geoscience and Remote Sensing* 49.2 (2011), pp. 717–725.

[57] Z. Li et al. "Swarm: mining relaxed temporal moving object clusters". In: *Proc. VLDB Endow.* 3.1-2 (2010), pp. 723–734.

[58] L. Libkin, J. L. Reutter, and D. Vrgoc. "Trial for RDF: adapting graph query languages for RDF data". In: *Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2013, New York, NY, USA - June 22 - 27, 2013*. 2013, pp. 201–212.

[59] Y. Lin et al. "Learning Entity and Relation Embeddings for Knowledge Graph Completion". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.* 2015, pp. 2181–2187.

[60] C. Low-Kam et al. "Mining Statistically Significant Sequential Patterns". In: *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*. 2013, pp. 488–497.

[61] F. P. S. Luus et al. "Multiview Deep Learning for Land-Use Classification". In: *IEEE Geosci. Remote Sensing Lett.* 12.12 (2015), pp. 2448–2452.

[62] Y. Ma et al. "Remote Sensing Big Data Computing". In: *Future Gener. Comput. Syst.* 51.C (Oct. 2015), pp. 47–60. ISSN: 0167-739X.

[63] Y. Ma et al. "Towards building a data-intensive index for big data computing - A case study of Remote Sensing data processing". In: *Inf. Sci.* 319 (2015), pp. 171–188.

[64] F. Mahdisoltani, J. Biega, and F. M. Suchanek. "YAGO3: A Knowledge Base from Multilingual Wikipedias". In: *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*. 2015.

[65] D. Marmanis et al. "Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks". In: *IEEE Geosci. Remote Sensing Lett.* 13.1 (2016), pp. 105–109.

[66] R. Meo, D.r Bachar, and D. Ienco. "LODE: A distance-based classifier built on ensembles of positive and negative observations". In: *Pattern Recognition* 45.4 (2012), pp. 1409–1425.

[67] A. Morales-González et al. "A new proposal for graph-based image classification using frequent approximate subgraphs". In: *Pattern Recognition* 47.1 (2014), pp. 169–177.

[68] G. W. Mueller-Warrant et al. "Methods for improving accuracy and extending results beyond periods covered by traditional ground-truth in remote sensing classification of a complex landscape". In: *Int. J. Applied Earth Observation and Geoinformation* 38 (2015), pp. 115–128.

[69] R. Navigli and S. P. Ponzetto. "BabelNet: Building a Very Large Multilingual Semantic Network". In: *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*. 2010, pp. 216–225.

[70] M. Nickel et al. "A Review of Relational Machine Learning for Knowledge Graphs". In: *Proceedings of the IEEE* 104.1 (2016), pp. 11–33.

[71] E. E. Papalexakis, L. Akoglu, and D. Ienco. "Do more views of a graph help? Community detection and clustering in multi-graphs". In: *Proceedings of the 16th International Conference on Information Fusion, FUSION 2013, Istanbul, Turkey, July 9-12, 2013*. 2013, pp. 899–905.

[72] Y. Pei, L.-P. Liu, and X. Z. Fern. "Bayesian Active Clustering with Pairwise Constraints". In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I*. 2015, pp. 235–250.

[73] R. G. Pensa, D. Ienco, and R. Meo. "Hierarchical co-clustering: off-line and incremental approaches". In: *Data Min. Knowl. Discov.* 28.1 (2014), pp. 31–64.

[74] Y. Pitarch et al. "Spatio-temporal data classification through multidimensional sequential patterns: Application to crop mapping in complex landscape". In: *Eng. Appl. of AI* 37 (2015), pp. 91–102.

[75] D. Ienco R. Bourqui A. Sallaberry and P. Poncelet. "Multilayer Graph Edge Bundling". In: *IEEE Pacific Visualization*. 2016, pp. 1–5.

[76] D. Redondo et al. "Layer-Centered Approach for Multigraphs Visualization". In: *19th International Conference on Information Visualisation, IV 2015, Barcelona, Spain, July 22-24, 2015*. 2015, pp. 50–55.

[77] M. Riondato and E. Upfal. "Mining Frequent Itemsets through Progressive Sampling with Rademacher Averages". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*. 2015, pp. 1005–1014.

[78] S. Romeo, D. Ienco, and A. Tagarelli. "Knowledge-Based Representation for Transductive Multilingual Document Classification". In: *37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*. 2015, pp. 92–103.

[79] S. Romeo, A. Tagarelli, and D. Ienco. "Semantic-Based Multilingual Document Clustering via Tensor Modeling". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. 2014, pp. 600–609.

[80] T. K. Saha and M. Al Hasan. "FS^3: A sampling based method for top-k frequent subgraph mining". In: *Statistical Analysis and Data Mining* 8.4 (2015), pp. 245–261.

[81] M. Seeland, A. Karwath, and S. Kramer. "A structural cluster kernel for learning on graphs". In: *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*. 2012, pp. 516–524.

[82] Arlei Silva, Wagner Meira Jr., and Mohammed J. Zaki. "Mining Attribute-structure Correlated Patterns in Large Attributed Graphs". In: *PVLDB* 5.5 (2012), pp. 466–477.

[83] C. A. R. de Sousa, S. O. Rezende, and G. E. A. P. A. Batista. "Influence of Graph Construction on Semi-supervised Learning". In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*. 2013, pp. 160–175.

[84] L. Tang, X. Wang, and H. Liu. "Community detection via heterogeneous interaction analysis". In: *Data Min. Knowl. Discov.* 25.1 (2012), pp. 1–33.

[85] N. Tatti. "Maximum entropy based significance of itemsets". In: *Knowl. Inf. Syst.* 17.1 (2008), pp. 57–77.

[86] N. Tatti and M. Mampaey. "Using background knowledge to rank itemsets". In: *Data Min. Knowl. Discov.* 21.2 (2010), pp. 293–309.

[87] S. Tong and D. Koller. "Support Vector Machine Active Learning with Application sto Text Classification". In: *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*. 2000, pp. 999–1006.

[88] L. Tsang and T. J. Jackson. "Satellite Remote Sensing Missions for Monitoring Water, Carbon, and Global Climate Change". In: *Proceedings of the IEEE* 98.5 (2010), pp. 645–648.

[89] D. Tuia et al. "Graph Matching for Adaptation in Remote Sensing". In: *IEEE T. Geoscience and Remote Sensing* 51.1 (2013), pp. 329–341.

[90] P. Poncelet V. Ingalalli D. Ienco. "On Querying Large Graphs with Multiple Relationships". In: *Conférence sur la Géstion de Données Principes, Technologies et Applications (BDA)*. 2015, pp. 1–12.

[91] P. Poncelet V. Ingalalli D. Ienco and S. Villata. "Querying RDF Data Using A Multigraph-basedApproach". In: *International Conference on Extending Database Technology (EDBT)*. 2016, pp. 1–12.

[92] V. Vapnik. *Statistical learning theory*. Wiley, 1998. ISBN: 978-0-471-03003-4.

[93] P. Vincent et al. "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion". In: *Journal of Machine Learning Research* 11 (2010), pp. 3371–3408.

[94] J. Vreeken, M. van Leeuwen, and A. Siebes. "Krimp: mining itemsets that compress". In: *Data Min. Knowl. Discov.* 23.1 (2011), pp. 169–214.

[95] J. Wang, J. Cheng, and A. Wai-Chee Fu. "Redundancy-aware maximal cliques". In: *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*. 2013, pp. 122–130.

[96] Y. Wang, E. P. Lim, and S. Y. Hwang. "Efficient mining of group patterns from user movement data". In: *Data Knowl. Eng.* 57.3 (2006), pp. 240–282. ISSN: 0169-023X.

[97] C.-Wei Wu et al. "Mining high utility episodes in complex event sequences". In: *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*. 2013, pp. 536–544.

[98] S. Xiong, J. Azimi, and X. Z. Fern. "Active Learning of Constraints for Semi-Supervised Clustering". In: *IEEE Trans. Knowl. Data Eng.* 26.1 (2014), pp. 43–54.

[99] C. Xue et al. "A spatiotemporal mining framework for abnormal association patterns in marine environments with a time series of remote sensing images". In: *Int. J. Applied Earth Observation and Geoinformation* 38 (2015), pp. 105–114.

[100] M. Yahya et al. "Relationship Queries on Extended Knowledge Graphs". In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*. 2016, pp. 605–614.

[101] X. Yan and J. Han. "gSpan: Graph-Based Substructure Pattern Mining". In: *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*. 2002, pp. 721–724.

[102] J. Soung Yoo and M. Bow. "Mining spatial colocation patterns: a different framework". In: *Data Min. Knowl. Discov.* 24.1 (2012), pp. 159–194.

[103] S. Zhang, J. Yang, and W. Jin. "SAPPER: Subgraph Indexing and Approximate Matching in Large Graphs". In: *PVLDB* 3.1 (2010), pp. 1185–1194.

[104] F. Zhu et al. "Mining Top-K Large Structural Patterns in a Massive Network". In: *PVLDB* 4.11 (2011), pp. 807–818.

Appendix D

Curriculum

Dino Ienco

CONTACT INFORMATION

Last Name: IENCO
First Name: DINO

Personal Address:
204, rue Dante Alighieri,
34790 Grabels, France
Professional Address:
UMR TETIS
500 rue Jean-François Breton,
34090 Montpellier
Tel: +33 (0)4.67.55.86.12
Mail: dino.ienco@irstea.fr
Web Site: https://sites.google.com/site/dinoienco/
Citizenship: Italy
Date of Birth: 27/06/1982

RESEARCH INTERESTS

Methodology

Data mining, Supervised/Semi-Supervised/Unsupervised Learning, Feature Selection, Anomaly Detection, Clustering, Associative Classification, Pattern mining, Graph Data Management

Data

Spatio-Temporal Data, Remote Sensing, Data Stream, Social Network, Text, RDF

ACTIVITY

Researcher (CR1) at **IRSTEA**, Montpellier, France
Associate Researcher in the Advanse Team, LIRMM, Montpellier, France
(from September 2011)

- Topic: Spatio-Temporal Data Mining

Post-Doc at **CEMAGREF**, Montpellier, France
(from February 2011 to August 2011)

- Topics: Geo-Spatial Data Mining, Sequence Mining

Post-Doc at **University of Torino**, Torino, Italy
Bioinformatic (from February 2010 until January 2011)

- BioBits project (Developing white and green biotechnologies by converging platforms from biology and information technology towards metagenomics)
- Topics: Data Management with open sources platform (GMOD, GBROWSE), Clustering approaches over genomes

EDUCATION

Ph.D. in Computer Science **University of Torino**, Torino, Italy (January 2010)

- Thesis Title: *Unsupervised approaches for the generation of structures on large data*
- Thesis Topic: Data Mining techniques to exploit hierarchical information from data
- Supervisor: Prof. Rosa Meo

M.S. in Computer Science **University of Torino**, Torino, Italy (September 2006)

- Thesis Topic: Data Mining application with emphasis on inductive database for sequence mining
- Supervisor: Prof. Marco Botta

Invited Talks

- Tri-National Scientific Workshop Climate change: Observation, Analysis and Health Indonesia Thailand France *October 2015*
  + Talk: *Detecting spatiotemporal dynamics in satellite remote sensing time series: methodological approach combining OBIA and data mining techniques*
- III School on Machine Learning and Knowledge Discovery in Databases - Sao Carlos - Brazil *October 2014*
  + Invited Lecture on *Active Learning, From Pool-Based to Stream Setting*

Visiting Activities

- Invited Visiting Researcher at USP, Sao Paolo, Brazil *April 2016* (Duration: One Month)
  + Anomaly detection in heterogenous data
- Invited Visiting Researcher at CNR, Milano, Italy *June 2014* (Duration: Two Weeks)
  + Spatio-Temporal Fuzzy Clustering
- Invited Visiting Researcher at the University of Waikato, Hamilton, New Zealand *February - April 2013* (Duration: Three Months)
  + Active Learning for data stream
  + Concept drift and change detection in categorical stream data
- Internship at Yahoo Research Lab!, Barcelona, Spain *November 2009 - January 2010* (Duration: Three Months)
  + Social network analysis on the Meme Yahoo data

Scientific Collaborations

- **International**:
  - *University of Torino* (Dr. Ruggero Pensa and Prof. Rosa Meo) 2007 - Ongoing.
    Study and development of new distance-based data mining approaches to deal with both categorical and textual data. In these scenarios distance computation can be negatively influenced by high-dimensionality (textual data [Ienco13j], [Pensa14j]) or unclear distance metric (categorical data [Ienco12j] and [Ienco16aj]).
  - *Yahoo Research Lab.* (Researcher Francesco Bonchi) 2009 - 2011.
    This collaboration was devoted to study and evaluate how the information spreads on real world social media considering both topological and content information [Bonchi13j].
  - *University of Bari* (Prof. Donato Malerba) 2012 - Ongoing.
    Study and Developing data mining approaches to mine useful correlations over relational data with a particular emphasis on gradual patterns [Phan15].
  - *Stony Broke University* (Dr. Leman Akoglu) 2014
    This collaboration was focused on the inspection and implementation of new community detection methods in the context of complex graph such as multilayer network [Papalexakis13].
  - *University of Waikato* (Prof. Bernhard Pfahringer) 2013 - 2014
    Study and development new data stream classification algorithms with a particular emphasis on active learning strategies. Active learning is useful to reduce the costly and tedious task of data labeling needed to constitute the train data and, in the case of data stream, not always feasible [Ienco13b], [Ienco14a] and [Ienco14b].
  - *Center for National Research Italy - CNR* (Researcher Gloria Bordogna) 2014 - Ongoing.
    Design and development of new spatio-temporal clustering methods with a particular emphasis on approaches able to exploit domain knowledge specified as fuzzy constraints [Bordogna14] and [Arcaini16j].
  - *University of Calabria* (Dr. Andrea Tagarelli) 2014 - Ongoing.
    Study and development of semantic-based textual representation for multilingual document collection. We exploit multilingual knowledge-bases to model and represent documents written in different languages. This new representation allows standard data mining techniques to cope with the language heterogeneity in comparable corpora [Romeo14a], [Romeo14b] and [Romeo15].

- *University of Sao Paolo (USP)* (Prof. André Carvalho Ponce de Leon) 2014 - Ongoing.
  Design and develop new anomaly detection methods especially tailored to deal with data represented by mixed attributes (numerical and categorical).

Supervised Students

- **PhD Students**
  - *Hai Phan Nhat* (percentage of supervision 40%) - Mining Object Movement Patterns from Trajectory Data, PhD founded by CNRS/Irstea, University of Montpellier 2 **Defended October 2013**. Co-Supervised with Dr. Maguelonne Teisseire (Irstea) and Prof. Pascal Poncelet (LIRMM). **Current situation** : post-doc at Oregon University, USA.
  - *Salvatore Romeo* (percentage of supervision 40%) - Multi-topic and Multilingual Document Clustering via Tensor Modeling, PhD founded by Italian Ministry of Education, University of Calabria **Defended March 2015**. Co-supervised with Dr. Andrea Tagarelli (Univ. of Calabria). **Current situation**: post-doc at QCRI, Qatar.
  - *Vijay Ingalalli* (percentage of supervision 50%) - Multigraph Query and Mining with Applications to Remote Sensing Images, PhD founded by NUMEV/Irstea, University of Montpellier, February 2014 Ongoing. Co-Supervised with Prof. Pascal Poncelet (LIRMM).
  - *Lionel Pribel* (percentage of supervision 30%): Detection of urban objects from heterogenous data sources, PhD founded by CIFRE (company Berger-Levrault), University of Montpellier, September 2015 - Ongoing. Co-Supervised with Dr. Marc Chaumont (LIRMM) and Dr. Gerard Subsol (LIRMM).
  - *Lynda Khiali* (percentage of supervision 50%): Mining spatio-temporal data from large volumes of satellite images, PhD founded by AVERROES, UMR TETIS, September 2015 - Ongoing. Co-Supervised with Dr. Maguelonne Teisseire (Irstea).
- **Post-docs**
  - *Fabio Güttler* (percentage of supervision 40%) - Approaches to mine and describe time series of remote sensing satellite images , Post-doc, founded by Equipex GEOSUD, June 2013 - December 2014 Co-Supervised with Dr. Maguelonne Teisseire (Irstea) and Prof. Pascal Poncelet (LIRMM). **Current situation**: post-doc at University of Strasbourg.
- **Master Students**
  - *Mykael Vigo* Twitter Event Detection and Modeling with TEWS, Master IPS (Informatique Pour les Sciences), University of Montpellier, 2015. Co-Supervised with Dr. Konstantin Todorov (LIRMM) and Prof. Zohra Bellahsene (LIRMM).
  - *Lionel Pibrel* Deep Learning approaches for Steganalysis, Master DECOL (Données, Connaissances et Langage Naturel), University of Montpellier, 2015. Co-Supervised with Dr. Marc Chaumont (LIRMM).
  - *Denis Redondo* Layer-Centered Approach for Multigraphs Visualization , Master Stic Santé, University of Montpellier, 2014. Co-Supervised with Dr. Arnaud Sallaberry (LIRMM) and Prof. Pascal Poncelet (LIRMM).
  - *Manel Achichi* Towards Linked Data Extraction From Tweets, Master DECOL (Donnes, Connaissances et Language Naturel), University of Montpellier, 2014. Co-Supervised with Dr. Konstantin Todorov (LIRMM) and Prof. Zohra Bellahsene (LIRMM).

Projects

- **DyNAmiTeF** CNES-Tosca project (2016-2017) DyNAmiTeF : DyNAmique des milieux NATurels par télédétection et Fouille de données (22k euros). Scientific Advisor.
- **JVWEB** (2013-2014) Aide à la faisabilité technologique (30k euros). Consulting about machine learning approaches for data stream. Co-supervised with Prof. Jerome Azé (LIRMM), Dr. Sandra Bringay (LIRMM) and Prof. Pascal Poncelet (LIRMM).
- **AE RMC** (2013-2015) Caractérisation des pressions agricoles par lutilisation de linformation spatialisée et de méthodes de fouilles de donnes - Modélisation pression / impacts pour la qualité des cours deau (67k euros - founded by the Water Agency). Scientific Co-Advisor with Dr. Maguelonne Teisseire (IRSTEA).
- **ANR FRESQUEAU** (2011-2015) Fouille de données pour l'évaluation et le suivi de la qualité hydrobiologique des cours d'eau (851K euros). Member of the project.

- **EQUIPEX GeoSud** (2011-2019) Infrastructure nationale dimagerie satellitaire pour la recherche sur lenvironnement et les territoires et ses applications la gestion et aux politiques publiques (21M euros). Participant through the supervision of a post-doc fellowship, Dr. Fabio Güttler, Co-Supervised with Prof. Pascal Poncelet (LIRMM) and Dr. Maguelonne Teisseire (IRSTEA).

Service and Volunteer Activities

International Conferences (Program Committee)
• **ACML** - Asian Conference on Machine Learning (2013, 2014, 2015)
• **ACM SAC** Track on Information Retrieval (2014, 2015)
• **ECML/PKDD** - European Conference on Machine Learning/Principle and Practice of Knowledge Discovery from Data (2013, 2014, 2015)
• **IJCAI** - International Joint Conference on Artificial Intelligence (2013)
• **ICDM** - IEEE International Conference on Data Mining (2013, 2014, 2015)
• **ECML/PKDD-DyNak** 2010 - Workshop on Dynamic Networks and Knowledge Discovery Discovery
• **IEEE DSAA** - Int. Conf. on Data Science and Adv. Analytics (2015)
• **ISCIS** - International Symposium on Computer and Information Sciences (2013)
• **KDIR** - International Conference on Knowledge Discovery and Information ReTrieval (2013)
• **NLDB** Natural Language and DataBase (2014)

International Journals (Reviewer)
• **Machine Learning Journal**, Springer (2014, 2015)
• **Data Mining and Knowledge Discovery**, Springer (2013, 2014, 2015)
• **Applied Soft Computing**, Elsevier (2014)
• **ACM Transaction on Intelligent Systems and Technology** (2013)
• **Information System**, Elsevier (2014)
• **ACM Transaction on Multimedia** (2013)
• **IDA Journal** - IOS Press (2013)
• **International Journal of Computers and Applications**, actapress (2013)
• **IEEE Transactions on Multimedia** (2013)
• **IJITDM** - International Journal of Information Technology And Decision Making (2013, 2014)
• **IEEE TKDE** Transaction on Knowledge and Data Engineering (2013, 2014, 2015)
• **ACM TKDD** Transaction on Knowledge Discovery from Data (2014)
• **Neurocomputing**, Elsevier (2014, 2015)
• **Theoretical Computer Science**, Elsevier (2014)
• **KAIS** - Knowledge and Information Systems, Springer (2016)

Phd Thesis Reviewer

- Reviewer of the PhD Thesis of Lucrezia Macchia: *Learning to Rank from Dynamic Network* (University of Bari, Italy), 2014

Workshop Organization

- *Artificial Intelligence meets the Web of Data* at European Conference on Artificial Intelligence (ECAI) 2012 co-chaired with Cristophe Gueret, Francois Scharffe and Serena Villata
- *Artificial Intelligence meets the Web of Data* at European Semantic Web Conference (ESWC) 2013 co-chaired with Cristophe Gueret, Francois Scharffe and Serena Villata
- *Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine)* at European Conference on Information Retrieval (ECIR) 2016 co-chaired with Salvatore Rome, Andrea Tagarelli, Mathieu Roche and Paolo Rosso

Teaching activity

Considering my teaching activities, I am involved in the Computer Science Master program at the University of Montpellier in DECOL and IPS programs. My implication in the Master programs are related to Data Mining and Machine Learning classes in which the base notion of these research fields are taught considering both lecture (CM, TD) and practice (TP) interventions.

Classes	University	Level	Year	Hours
Fouille de données	Univ. of Montpellier, France	Master 1 - IPS	2015/2016	12h
ECD	Univ. of Montpellier, France	Master 1 - DECOL	2015/2016	10h
ECA	Univ. of Montpellier, France	Master 2 - DECOL	2015/2016	15h
Fouille de données	Univ. of Montpellier, France	Master 1 - IPS	2014/2015	12h
ECD	Univ. of Montpellier, France	Master 1 - DECOL	2014/2015	19h
ECA	Univ. of Montpellier, France	Master 2 - DECOL	2014/2015	13h
Active learning	Univ. of Sao Paolo, Brazil	Summer School	2014	3h
ECD	Univ. of Montpellier, France	Master 1 IPS	2013/2014	3h
ECA	Univ. of Montpellier, France	Master 2 - DECOL	2013/2014	20h
ECD	Univ. of Montpellier, France	Master 1 - IPS	2012/2013	10h
ECD	Univ. of Montpellier, France	Master 1 - IPS	2011/2012	15h
BD	Univ. of Montpellier 2 , France	Licence 2	2011/2012	33h

Publications

JOURNALS

[**Ienco16aj**] D. Ienco and R.G. Pensa *Positive and unlabeled learning in categorical data* Neurocomputing - Accepted for Publication, pp. 1-24 (2016).

[**Ienco16bj**] D. Ienco, R.G. Pensa and R. Meo *A Semi-Supervised Approach to the Detection and Characterization of Outliers in Categorical Data* Transaction on Neural Network and Learning Systems - Accepted for Publication, pp. 1-13.

[**Güttler16j**] F. Güttler, D. Ienco, P. Poncelet and M. Teisseire *Combining Transductive and Active Learning to Improve Object-based Classification of Remote Sensing Images* Remote Sensing Letters - Accepted for Publication, pp. 1-10 (2016).

[**Arcaini16j**] P. Arcaini, G. Bordogna, D. Ienco and S. Sterlacchini *User-driven geo-temporal density-based exploration of periodic and not periodic events reported in social networks* Information Sciences - Accepted for Publication, pp. 1-33 (2016).

[**Berrahou15j**] S. L. Berrahou, N. Lalande, E. Serrano, G. Molla, L. Berti-Équille, S. Bimonte, S. Bringay, F. Cernesson, C. Grac, D. Ienco, F. Le Ber and M. Teisseire *A quality-aware spatial data warehouse for querying hydroecological data* Computers & Geosciences, 85: 126-135 (2015).

[**Pitarch15j**] Y. Pitarch, D. Ienco, E. Vintrou, A. Bégué, A. Laurent, P. Poncelet and M. Teisseire *Spatio-Temporal Data Classification through multi-dimensional sequential patterns: application to food risk analysis* Engineering Application of Artificial Intelligence (EAAI), 37: 91-102 (2015).

[**Egho14j**] E. Egho, N. Jay, D. Ienco, C. Raissi, P. Poncelet, M. Teisseire and A. Napoli *A contribution to the discovery of multidimensional patterns in healthcare trajectories.* Journal of Intelligent Information Systems (JIIS), 42(2): 283-305 (2014).

[**Pensa14j**] R.G. Pensa, D. Ienco and R. Meo *Hierarchical Co-Clustering: Off-line and Incremental Approaches* Data Mining and Knowledge Discovery Journal (DAMI), 28(1): 31-64 (2014).

[**Vintrou13j**] E. Vintrou, D. Ienco, A. Begue and M. Teisseire. *Data mining, a promising tool for large area cropland mapping* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6(5): 2132-2138 (2013).

[**Bonchi13j**] F. Bonchi, C. Castillo and D. Ienco *Meme Ranking to Maximize Posts Virality in Microblogging Platforms* Journal of Intelligent Information System (JIIS), 40(2): 211-239 (2013).

[**Ienco13j**] D. Ienco, C. Robardet, R. G. Pensa and R. Meo *Parameter-Less Co-Clustering for Star-Structured Heterogeneous Data* Data Mining and Knowledge Discovery Journal (DAMI), 26(2): 217-254 (2013).

[**Ienco12j**] D. Ienco, R.G. Pensa and R. Meo. *From Context to Distance: Learning Dissimilarity for Categorical Data Clustering* ACM Transaction on Knowledge Discovery from Data (TKDD), 6(1): 1-29 (2012).

[**Meo12j**] R. Meo, D. Bachar and D. Ienco *A distance-based classifier built on ensembles of positive and negative observations* Pattern Recognition, 45(4): 1409-1425 (2012).

[**Bonfante11j**] P. Bonfante , F. Cordero , S. Ghignone , D. Ienco , L. Lanfranco, G. Leonardi, R. Meo, S. Montani, L. Roversi, and A. Visconti *A Modular Database Architecture Enabled to Comparative Sequence Analysis* LNCS Transactions on Large-Scale Data and Knowledge-Centered Systems, Springer, 124-147 (2011).

CHAPTER IN BOOKS

[**Visconti12**] A. Visconti, F. Cordero, D. Ienco and R.G. Pensa. *Coclustering under Gene Ontology derived Constraints for Pathway Identification.* Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data, M. Elloumi and A.Y. Zomaya (Eds). Wiley, USA. pp. 625-642, 2012.

[**Meo09**] R. Meo and D. Ienco *Replacing Support in Association Rule Mining* invited chapter in Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection, Yun Sing and Nathan Rountree (ed.), Advances in Data Warehousing and Mining Book Series, IGI Global publisher, 2009. ISBN: 1935-2646.

INTERNATIONAL CONFERENCES

[**Bourqui16**] R. Bourqui, A. Sallaberry, D. Ienco and P. Poncelet *Multilayer Graph Edge Bundling*, IEEE Pacific Visualization 2016 (PacificVis16), pp. 1-5.

[**Ingalalli16**] V. Ingalalli, D. Ienco, P. Poncelet and S. Villata *Querying RDF Data Using A Multigraph-based Approach*, International Conference on Extending Database Technology (EDBT16), pp. 1-12.

[**Redondo15**] D. Redondo, A. Sallaberry, D. Ienco, F. Zaidi and P. Poncelet *Layer-Centered Approach for Multigraphs Visualization*, Information Visualisation Theory And Practice (IV15), pp. 50-55.

[**Phan15**] H. Phan Nhat, D. Ienco, D. Malerba, P. Poncelet and M. Teisseire. *Mining Multi-Relational Gradual Patterns*, Siam on Data Mining (SDM15), pp. 846-854

[**Romeo15**] S. Romeo, D. Ienco and A. Tagarelli. *Knowledge-based Representation for Transductive Multilingual Document Classification*, European Conference on Information Retrieval (ECIR15), pp. 92-103.

[**Romeo14a**] S. Romeo, A. Tagarelli and D. Ienco. *Clustering View-Segmented Documents via Tensor Modeling*, International Symposium on Methodologies for Intelligent Systems (ISMIS14), pp. 385-394.

[**Romeo14b**] S. Romeo, A. Tagarelli and D. Ienco. *Semantic-Based Multilingual Document Clustering via Tensor Modeling*, Conference on Empirical Methods in Natural Language Processing (EMNLP14), pp. 600-609.

[**Bordogna14**] D. Ienco and G. Bordogna *Fuzzy Core DBScan Clustering Algorithm*, Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU14), pp. 100-109.

[**Güttler14a**] F. Güttler, Dino Ienco, Maguelonne Teisseire, Jordi Nin, Pascal Poncelet *Towards the Use of Sequential Patterns for Detection and Characterization of Natural and Agricultural Areas*, Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU14), pp. 97-106.

[**Güttler14b**] F. Güttler, S. Alleaume, C. Corbane, D. Ienco, J. Nin, P. Poncelet and M. Teisseire. *Exploring high repetitivity remote sensing time series for mapping and monitoring natural habitats - a new approach combining OBIA with K-Partite Graph*, IEEE International Geoscience and Remote Sensing Symposium (IGARSS14), pp. 3930-3933.

[**Ienco14a**] D. Ienco, A. Bifet, B. Pfahringer and P. Poncelet *Change Detection in Categorical Evolving Data Streams*, ACM Symposium of Applied Computing (SAC14), pp. 792-797.

[**Ienco13a**] D. Ienco, Y. Pitarch, P. Poncelet and M. Teisseire *Knowledge-free Table Summarization*, International Conference on Data Warehousing and Knowledge Discovery (DAWAK13), pp. 122-133.

[**Ienco13b**] D. Ienco, A. Bifet, I. Zliobaitè and B. Pfahringer *Clustering based Active Learning for Evolving Data Streams*, Discovery Science (DS13), pp. 79-93.

[**Papalexakis13**] E. E. Papalexakis, L. Akoglu and D. Ienco. *Do more views of a graph help? Community Detection and Clustering in Multi-Graphs*, IEEE International Conference on Information Fusion (FUSION13), pp. 899-905.

[**Phan13**] H. Phan Nhat, D. Ienco, P. Poncelet and M. Teisseire. *Mining Representative Movement Patterns through Compression*, Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD13), pp. 314-326.

[**Phan12a**] H. Phan Nhat, D. Ienco, P. Poncelet and M. Teisseire. *Mining Time Relaxed Gradual Moving Object Clusters*, ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS12), pp. 478-481.

[**Phan12b**] H. Phan Nhat, D. Ienco, P. Poncelet and M. Teisseire. *Mining Fuzzy Moving Object Clusters*, International Conference on Advanced Data Mining and Applications (ADMA12), pp. 100-114.

[**Loglisci12**] C. Loglisci, D. Ienco, M. Roche, M. Teisseire and D. Malerba *An Unsupervised Framework for Topological Relations Extraction from Geographic Documents*, International Conference on Database and Expert Systems Applications (DEXA12), pp. 48-55.

[**Ienco12**] D. Ienco, Y. Pitarch, P. Poncelet and M. Teisseire *Towards an Automatic Construction of Contextual Attribute-Value Taxonomies*, ACM Symposium on Applied Computing, Data Mining Track (SAC12), pp. 113-118.

[**Spinella11**] S. Spinella, E. Sciacca, D. Ienco and P. Giannini *Annotated Stochastic Context Free Grammars for Analysis and Synthesis of Proteins*, Eur. Conf. on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EVOBIO11), pp. 77-88.

[**Boella11**] G. Boella, S. Colombo Tosatto, A. d'Avila Garcez, V. Genovese, D. Ienco and L. van der Torre *Neural Symbolic Systems for Normative Agents* , International Conference on Autonomous Agents and Multiagent Systems (AAMAS11), pp. 1203-1204.

[**Cordero09**] F. Cordero, R.G. Pensa, A. Visconti, D. Ienco, M. Botta. *Ontology-driven Coclustering of Gene Expression Data*, Conference of the Italian Association for Artificial Intelligence (AI*IA09), pp. 426-435.

[**Ienco09a**] D. Ienco, R.G. Pensa and R. Meo. *Parameter-free Hierarchical Co-Clustering by N-Ary Splits*, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD09), pp. 580-595.

[**Ienco09b**] D. Ienco, R.G. Pensa and R. Meo. *Context-based Distance Learning for Categorical Data Clustering*, Symposium on Intelligent Data Analysis (IDA09), pp. 83-94.

[**Ienco08a**] D. Ienco, R. Meo. *Towards Automatic Construction of Conceptual Taxonomies*, International Conference on Data Warehousing and Knowledge Discovery (DAWAK08), pp. 327-336.

[**Ienco08b**] D. Ienco, S. Villata and C. Bosco. *Subcategorization frame extraction for Italian*, Language Resources Evaluation Conference (LREC08).

[**Ienco08c**] D. Ienco, R. Meo. *Exploration and Reduction of the Feature Space by Hierarchical Clustering*, SIAM International Conference on Data Mining (SDM08), pp. 577-587.

NATIONAL CONFERENCES

[**Ingalalli15nc**] V. Ingalalli, D. Ienco and P. Poncelet *SuMGRA: On Querying Large Graphs with Multiple Relationships* Conférence sur la Gestion de Données Principes, Technologies et Applications (BDA15).

[**Ienco14nc**] D. Ienco, A. Bifet, B. Pfahringer and P. Poncelet *Détection de changements dans des flots de données qualitatives* Conference Internationale Francophone sur l'Extraction et la Gestion de Connaissance (EGC14).

[**Ienco12nc**] D. Ienco, Y. Pitarch, P. Poncelet and M. Teisseire *Vers une methode automatique pour la construction de hirarchies contextuelles* Conference Internationale Francophone sur l'Extraction et la Gestion de Connaissance (EGC12).

[**Ienco09nc**] D. Ienco, R. Meo. *Distance based Clustering for Categorical Data* Italian Symposium on Advanced Database Systems (SEBD09).

[**Ienco08anc**] D. Ienco, R. Meo. *Clustering the Feature Space* Italian Symposium on Advanced Database Systems (SEBD08).

[**Ienco08bnc**] D. Ienco, R. Meo, M. Botta. *Using PageRank in Feature Selection* Italian Symposium on Advanced Database Systems (SEBD08).

INTERNATIONAL WORKSHOPS AND DEMOS

[**Vigo15**] M. Vigo, Z. Bellahsene, D. Ienco and K. Todorov Twitter *Event Detection and Modeling with TEWS* International Semantic Web Conference - Demo (ISWC15).

[**Pensa14**] R.G. Pensa and D. Ienco *Learning from Categorical Attribute Relationships for Positive-Unlabeled Classification* European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (RL14@ECML-PKDD14).

[**Ienco14b**] D. Ienco, I. Zliobaite and B. Pfahringer *High density-focused uncertainty sampling for active learning over evolving stream data* ACM KDD Conference (BigMine14@KDD14)

[**Loglisci12**] C. Loglisci, D. Ienco, M. Roche, M. Teisseire and D. Malerba *Toward Geographic Information Harvesting: Extraction of Spatial Relational Facts from Web Documents* ICDM IEEE Workshop Spatial and Spatio-Temporal Data Mining (SSTDM12@ICDM12).

[**Egho12**] E. Egho, D. Ienco, N. Jay, A. Napoli, P. Poncelet, C. Quantin, C. Raissi and M. Teisseire *Healthcare Trajectory Mining by Combining Multi-dimensional Component and Itemsets* European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (NFMCP@ECML-PKDD12).

[**Phan12**] N. H. Phan, D. Ienco, P. Poncelet, and M. Teisseire *Extracting Trajectories through an Efficient and Unifying Spatio-Temporal Patten Mining System* European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - Demo (ECML-PKDD12).

[**Ienco10**] D. Ienco, F. Bonchi, C. Castillo. *The Meme Ranking Problem: Maximizing Microblogging Virality* ICDM IEEE Workshop on Social Interactions Analysis and Service Providers - (SIASP10@ICDM10).

Patents

- R. Cezar, D. Ienco, A. Mas, F. Masseglia, P. Poncelet, P. Pudlo, E. Szekely, M. Teisseire and J.P. Vendrell. "(WO2014118343) PROCESS FOR IDENTIFYING RARE EVENTS". International Patent, PCT/EP2014/051963, 2014.