13

Mining Spatio-Temporal Datasets: Relevance, Challenges and Current Research Directions

M-Tahar Kechadi[1], Michela Bertolotto[1],
Filomena Ferrucci[2] and Sergio Di Martino[2]
[1]*School of Computer Science and Informatics, Univeristy College Dublin*
[2]*Dipartimento di Matematica e Informatica Università di Salerno,*
[1]*Ireland*
[2]*Italy*

1. Introduction

Spatio-temporal data usually records the states over time of an object, an event or a position in space. Spatio-temporal data can be found in several application fields, such as traffic management, environment monitoring, weather forecast, etc. In the past, huge effort was devoted to spatial data representation and manipulation with particular focus on its visualisation. More recently, the interest of many users has shifted from static views of geospatial phenomena, which capture its "spatiality" only, to more advanced means of discovering dynamic relationships among the patterns and events contained in the data as well as understanding the changes occurring in spatial data over time.

Spatio-temporal datasets present several characteristics that distinguish them from other datasets. Usually, they carry distance and/or topological information, organised as multidimensional spatial and temporal indexing structures. The access to these structures is done through special methods, which generally require spatial and temporal knowledge representation, geometric and temporal computation, as well as spatial and temporal reasoning. Until recently, the research in spatial and temporal data handling has been mostly done separately. The research in the spatial domain has focussed on supporting the modelling and querying along spatial dimensions of objects/patterns in the datasets. On the other hand, the research in the temporal domain has focussed on extending the knowledge about the current state of the system governed by the temporal data. However, spatial and temporal aspects of the same data should be studied in conjunction as they are often closely related and models that integrate the two can be beneficial to many important applications.

Indeed the amount of available spatio-temporal datasets is growing at exponential speed and it is becoming impossible for humans to effectively analyse and process. Suitable techniques that incorporate human expertise are required. Data mining techniques have been identified as effective in several application domains. In this chapter we discuss the application of data mining techniques to effectively analyse very large spatio-temporal datasets.

Spatio-temporal data mining is an emerging field that encompasses techniques for discovering useful spatial and temporal relationships or patterns that are not explicitly stored in spatio-temporal datasets. Usually these techniques have to deal with complex objects with spatial, temporal and other attributes. Both spatial and temporal dimensions add substantial complexity to the data mining process. Following the above mentioned

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca,
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

separation traditionally applied to the analysis of spatial and temporal dimensions, spatial data mining (SDM) and temporal data mining (TDM) have received much attention independently within the KDD (Knowledge Discovery in Databases) research community. Motivations that have kept these fields separate include:

- **Complexity**: the investigation of spatio-temporal relations complicates the data mining process. The existing spatial or temporal techniques are not suitable for such additional constraints in terms of data types, data representation and structures. Therefore, the exploration of efficient techniques for spatio-temporal data becomes a necessity.
- **Data Models**: the absence of efficient space-time data models makes it difficult for the development of data mining techniques that deal with space and time simultaneously. Most existing data mining techniques applied to spatio-temporal datasets use very simple representations of the objects and their relationships, which usually based on spatial or temporal, but not both. The first model that integrates the time and space was proposed in (Peuquet et al., 1995). Since 1995 other models have been proposed in (Yuan, 1997). However, these models are too generic to handle special cases and implicit complex relationships, as it is often the case in spatial and temporal real-world data mining applications.
- **Application Domain**: An application is usually classified as spatial or temporal depending on the target problem at hand and how the datasets are collected. For instance, consider data describing three regions in Ireland, collected in three successive years. Each region is divided into grids of cells of the same dimensions. The goal is to predict the spatial distribution of agriculturally beneficial herbs e.g. Birdsfoot trefoil. The collected data, which consists of a certain number of attributes (soil pH, N, P, K status, herb abundance, distance from field boundary structure, e.g., hedgerow, etc.), chosen according to the domain knowledge, records the values of these attributes in each cell of the three regions. In this example, even though the overall goal is prediction, the time dimension is not taken into account in the model, as there are only three timestamps. Therefore, more effort is put into the modelling of the space dimensions and the data is recorded accordingly.

The development of efficient techniques for combined spatial and temporal data mining is an open and challenging issue within the research community. In this chapter we investigate this topic and discuss current trends in the area. In particular, given the visual aspect of spatial data, we discuss how visualisation techniques have been applied to support the spatio-temporal data analysis and mining process.

The remained of the chapter is organised as follows. In Section 2 and 3 we explain spatial and temporal data mining, respectively, the challenges of the corresponding mining tasks, review existing approaches, and discuss the main research directions. Section 4 is dedicated to on-going efforts to develop data mining techniques that take into account both spatial and temporal aspects of data. In Section 5 visual techniques applied in support of the mining process are discussed, while in Section 6 we present some innovative techniques for effectively mining very large spatio-temporal datasets as part of our work in this area. Finally, Section 7 concludes summarising new research directions and challenges.

2. Spatial data mining

Spatial data is characterised by spatial location attributes or dimensions. These spatial attributes are usually stored as coordinates and topology. Moreover spatial data contains

also non-spatial attributes. The spatial dimensions and their features add another level of complexity compared to relational data in terms of both the mining efficiency and the complexity of possible patterns that can be extracted from spatial datasets (Hinnenburg and Keim, 1998), (Roddick and Lees, 2001), and (Shekhar et al., 2001). The main reasons include:

1. The attributes of neighbouring patterns may have significant influence on a pattern and should also be considered.
2. Classical analytical methods for spatial data were introduced when the technology for collecting such data was very expensive and not powerful and, consequently, the available datasets were small and few. Nowadays, massive amounts of spatial data are collected daily at relatively low cost and classical methods cannot cope with them.
3. While non-spatial datasets are mainly composed of discrete objects stored in databases with well-defined relationships, spatial objects are embedded in a continuous space and characterised by implicit topological, distance and directional relationships (Bedard et al., 2001).

Much research has been dedicated to addressing these issues. Techniques that have been proposed include spatial classification, spatial association, spatial clustering, spatial outlier analysis and spatial prediction. Approaches based on spatial classification group spatial objects into categories based on distance, direction and connectivity relationships among them. For example, (Koperski et al., 1998) introduced spatial buffers to classify objects based on attribute similarity measures and proximity. (Ester et al., 2001) generalised this approach to other different spatial relationships. Spatial association rules are rules that rely on spatial predicates. The work by (Koperski et al., 1996) and (Yoo et al., 2005) provide examples of methods based on the spatial association approach. Methods for spatial clustering borrow heuristics proposed for general clustering algorithms to define meaningful groupings within the input data. Traditional techniques such as k-means and expectation maximisation can take distance relationships into account and can therefore be applied to spatial data in a straightforward way (Han et al., 2001). Alternative approaches include: hierarchical, grid-based, constraint-based, and density-based methods. Spatial outliers are spatial objects whose non-spatial attribute values are inconsistent with other objects, which are in the same local (spatial) neighbourhood. (Shekar et al., 2003) proposed a method for detecting spatial outliers efficiently. (Ng, 2001) adopted a similar approach to identify unusual trajectories based on endpoints, speed and geometry of the trajectories.

Prediction methods combine classification with inductive learning and/or artificial neural networks approaches to extract information from different types of spatial datasets. Examples include techniques applied to topographic maps (Malerba et al., 2001), soil-landscape maps (Qi & Zhou, 2003) and remotely sensed imagery (Gopal et al., 2001).

As previously mentioned, many spatial datasets also have an associated temporal dimension and, therefore, are referred to as spatio-temporal. Integrating time introduces additional complexity to the mining and knowledge discovery process. Simply adding another dimension does not represent an effective solution. Indeed, time and space have different semantics. We will develop spatio-temporal data mining in section 4. The next section is devoted to temporal data mining.

3. Temporal data mining

In general, temporal data mining techniques are designed for mining large sequential datasets. A set of data is said to be sequential if its data is ordered with respect to some

index, such as time. For instance, time series datasets are a common class of sequential data, which is recorded according to the index time. Moreover, there are many other sequential data that are not depending on the time. These include protein and gene sequences, Web click streams, sequences of moves in games such as Go and chess, alarms generation in telecommunication networks, etc. The overall goal is to discover sequential relationships or patterns that are implicitly present in the data. These sequences (sequential patterns) can be very useful for many purposes, for example, prediction of future event sequences. Consider a telecommunication company, the sequential patterns can be used for customer churn, marketing new products, pricing, and many others tasks.

The time series problems can behave in four different way: linearly, stationary, periodically, or randomly. A time series application can be represented as a linear problem when the future observation can be a linear function of the past observations. A time series can be stationary when it has a constant mean and variance. For non-stationary time series, the future observations cannot be foreseen, as they are very difficult to model. Time series can be periodic; displaying dominant periodic components with regular periodic variations. Finally, time series can be a random noise problem, which means that there can be a random noise included in some parts or the entire frequency spectrum of the time series.

There are several techniques implemented to handle time series applications. For instance, the simple moving average and exponential moving average techniques are used to deal with linear and stationary time series problems, the simple regression methods, auto regressive and auto regressive average deal with non-stationary time series problems, and decomposition methods deal with seasonal time series problems. One difficulty with regression techniques is that the correlation between the component variables, which affects the observation demand, is not stationary but depends on spatial-temporal attributes. Therefore, they are not capable of tackling this chronological disparity. Moreover, most of these techniques are problem-dependent, which means that while they return reasonably good solutions to the application at hand, they are not suitable for other types of temporal datasets.

3.1 Output patterns

Temporal data mining have been heavily investigated mainly for **prediction**, **classification**, and **clustering**. The **prediction** task in the time-series applications deals with forecasting future values of the time series based on its current and past values. Usually, prediction requires an efficient predictive model for the data. By model we mean an abstract representation of the data. For example neural network or Markov models have been widely used for prediction in sequence classification and forecasting applications (Rashid et al., 2006).

Classification assumes that some classes or categories have already been predefined. The main objective is to automatically identify for each input sequence its corresponding class or category. There are several sequence classification applications. These include handwriting recognition (Fitzgerald et al. 2004), gene sequences, speech recognition, currency exchange rates, stock market prices, etc. The temporal data mining technique for classification task are divided into two categories (Laxman et al., 2006): *model-based* methods and *pattern-based* methods. Pattern-based methods use a database of prototype feature sequences. Each class is represented by a set of prototype feature sequences (Ewens et al., 2001). For any given input sequence, the classifier searches over all prototypes looking for the closest (or most

similar) to the features of the new input. The model-based methods are techniques that use some powerful existing models such as Hidden Markov models, neural networks, support vector machines, etc. Usually these models consist of two phases: learning phase and testing phase. During the learning phase the model is trained on examples of each pattern class. The model assigns to the new pattern a class that contains the most likely pattern to generate it.
Unlike classification, **clustering** does not assume the class labels. Clustering groups the sequences together based on their similarity. Basically, sequences that are similar are grouped together and those that are dissimilar are assigned to different groups or clusters. Clustering is particularly interesting as it provides a dynamic mechanism for finding some structures (or clusters) in large datasets without any assumption about the number of the groups (clusters) in advance.

3.2 An example

In this section we consider the task of electric load forecasting for large and complex buildings. A correct estimation of the energy in this case is crucial as it can result in substantial savings for a power system. Once modelled correctly, it allows planning and/or designing of new future plants, providing security and reliability, and savings in the operational cost of a power system. It is apparent that there are relationships between the energy load and factors affecting it, yet these relations have not been clearly defined and understood. This problem is complex and, in order to learn something from this historical dataset, a very good data model and attributes' selection are required. For many years this application has been classified as one of the important class of time-series. We will see later in this chapter that this can also be modelled as a spatio-temporal data mining application when it involves various regions/locations of the world.
In this section, we consider the datasets that reflects the behaviour of the electricity supply in the Republic of Ireland. The main recorded attributes are the load, temperature, cloud rate, wind speed and humidity at 15mn intervals. This is a typical time-series application; each sequence of recorded data represents a value of a particular feature; these features are observed at different time intervals.
The choice of a data model is a very crucial and complex task in data mining. Not only should the model represent the data precisely but it should also be appropriate for the mining technique used. For instance the data inputs of a neural network technique are different from the inputs of a support vector machine or a hidden Markov model. We usually divide the dimensions into two categories: primary and secondary. Primary dimensions are the main dimensions that characterise the data itself. The secondary dimensions are informative but they can play a huge role when associated with the inputs of a given mining technique. The difficulty here is that there is no general rule for how to select appropriate secondary dimensions (inputs) for a given mining process. The selection depends largely on the experience and the expertise of the user within that specific domain or application.
In the case of energy load forecasting, we can define four types of attributes:

1. **The time and seasonal attributes**: As for any time series application, the time is one of the most important dimensions in this task. However, the time on its own is not enough. The variation in daily load profile is mostly affected by the localised weather effects and seasonal changes, which introduces weather patterns. Therefore, it is

imperative to include time information such as the time of the day, the day of the week, and the season of the year in order to model appropriately the forecasting behaviour.

2. **Direct and indirect weather attributes**: Direct weather attributes play a key role in the energy load model. These attributes include temperature, cloud rate, wind speed, wind direction, humidity, rainfall, etc. Along with time dimensions, all these attributes constitute principal dimensions of the model. Indirect or secondary attributes provide extra information about the application. These include relative change in temperature and relative change in load linked with the day, month or season of the year.
3. **The status of the day**: The load consumption depends on other factors such as special days; weekends, holidays, and other special events. Therefore these external factors should be included in the model. Some different behaviour was also noticed on the days before and after weekends and holidays. They are also treated as another different status of the day.
4. **Historical absolute/relative change in the load**: With this attribute one wants to model relationships between the attributes of a sequence pattern. In other words, one wants to identify the relationship between two consecutive days or the change in the load between two consecutive days. This notion should be extended to other attributes such as cloud rate, humidity, wind speed, temperature, etc.

Generally, some of the attributes involved have to be normalised in order to avoid errors in prediction due to the increase in electricity load consumption that is necessary for the economy growth. Moreover, the economy growth rate is needed for the final calculations of the predicted values; this should be part of the data pre-processing phase. As mentioned above, there are some attributes, which are not primary but necessary for a time-series application, such as energy load forecasting. These are called secondary or endogenous attributes. Example of daily temperature collected for January 1997 and 1998, shown if Figure (1) illustrates the difference between primary (exogenous) attribute and its corresponding secondary *endogenous) attribute. For instance in Figure 1a, it is difficult to see the patterns between the temperature in January 1997 and 1998, while in Figure 1b we can easily see that the change in temperature through the two months presents some patterns. The goal is to exploit theses changes in temperature in order to extract useful sequences (patterns) and use them for future predictions.

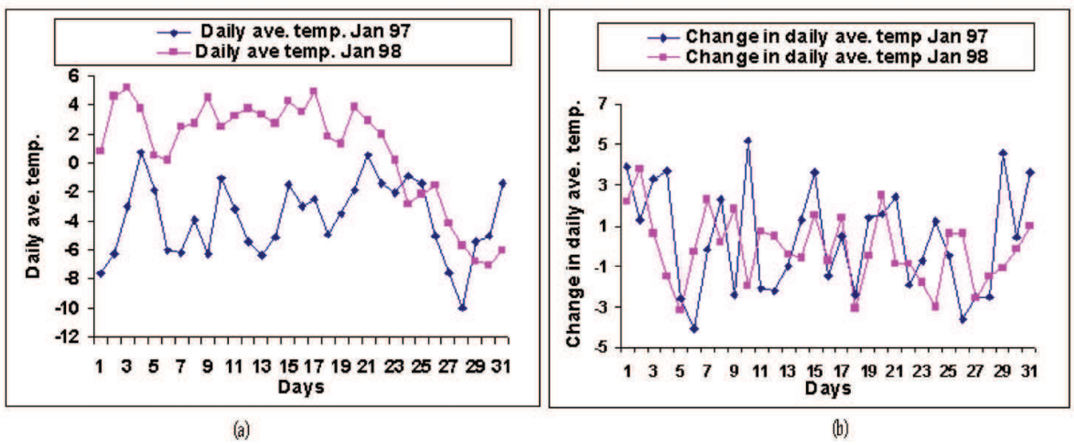

Fig. 1. (a) Daily Average Temperature (DAT). (b) the difference of DAT between two consecutive days.

Different data mining techniques have been used for this application. These techniques are mainly based on recurrent neural networks (RNN) (Elman, 1990) and Kohonen networks (Kohonen, 1995), Hidden Markov Models (HMM) (Picone, 1990), and Support Vector Machines (SVM) (Cortes et al., 1995). Essentially, we have developed several different variants of these techniques. In the case of SRNs, we proposed innovative architecture that can cope with time-series characteristics (Huang et al., 2006, Huang et al. 2005, Huang et al. 2004). It was shown that this new network architecture models more accurately the energy load forecasting and is about 20% more efficient than SVM or HMM based techniques (Tarik et al., 2006). We have also developed another hybrid technique that combines Kohonen networks with traditional data mining technique (k-means), the results were very promising, approaching 98,6% prediction accuracy (Gleeson et al., 2006). This motivated us to develop another hybrid technique based on the ensemble networks principal. The first version of this approach consisted of the combination of the previous developed techniques. While it was quite tricky to find a good system for combining the results of each technique, the overall results were better than the average of the results of the individual solutions. We are currently exploring a different approach of getting the benefit of each approach by building some interactions between them. These interactions are represented as some intermediate rules, allowing some of the techniques to change the its behaviour based on this new input.

4. Spatio-temporal mining

Spatio-temporal Data Mining is an emerging research area (Roddick and Lees, 2001), encompassing a set of exploratory, statistical and computational machine learning approaches for analysing very large spatial and spatio-temporal datasets. Presently, several open issues can be identified in this research field ranging from the definition of suitable mining techniques able to deal with spatio-temporal information to the development of effective methods to analyse the produced results.

In spatio-temporal datasets an event is described as a spatial and temporal phenomenon, as we consider that it happens at a certain time t and at a location x. Examples of event types include hurricanes, tornados, road traffic jam, road accidents, etc. In real world many of these events interact with each other and exhibit spatial and temporal patterns, which may help us understand the physical phenomenon behind them. Therefore, it is very important to identify efficiently the spatial and temporal features of these events and their relationships from large spatio-temporal datasets of a given application domain.

Spatio-temporal data mining presents many challenges to both researchers and users communities (Compieta et al., 2007). For researchers, the key challenges are to develop efficient and general methods that can support complex spatio-temporal data types structures as well as the scalability issue as the amount of currently collected data increases at exponential rates. The relationships between the spatial and temporal aspects of the data are often not clear or well defined. Providing means of extracting or defining such relationships is difficult and it is one the main objectives of data mining approaches in this area. Finally, the granularity levels have direct impact on these relationships between spatial and temporal features, so deciding or determining which level(s) is(are) more appropriate to investigate is a very challenging issue. For the users, even though they might have a deep understanding of the application at hand, the key issue is have to find a proper model that supports the given data mining approach. This step is not easy as there are no general rules

of how to build such a model. This type of experimentations with the data and expertise are still part of the current research efforts.

While these challenges remain hot topics in the area, there has been a good progress, in data preparation, models, dependency analysis, etc., in recent years. Based on the models that have been developed for spatial and temporal data of real-world applications, we can classify these applications into four categories (Yao, 2003): 1) Applications where the time is not part of the recorded data (may be not important). In this case there is no way of extracting patterns from the data that include the time dimension. These include all purely spatial data mining applications. 2) Applications where the data is recorded as ordered sequences of events according to a specific relation such as before and after, time-stamp, etc. 3) Applications where the data is recorded at regular intervals, and finally 4) applications where the dimension time is fully integrated in the recorded data. The application example given in section 3.2 can be classified into the third category.

We can define spatio-temporal data as a set of spatio-temporal sequences, S. Each element of the sequence is represented by its spatial and temporal attributes (x_1, x_2, …, x_n, t), where x_i, $1 \leq I \leq n$, is a spatial attribute and t a temporal attribute. For sake of clarity, we consider only one principal temporal attribute and we are not mentioning here non-spatial attributes. The goal is to study the behaviour of some objects (events) in the space or through the time. An example would be the study of the movement of a Hurricane. In this example, one should define what is an object in a Hurricane and then track its movement in the space. Moreover, we need to take into account the surrounding of an object of a Hurricane; therefore more than one object should be tracked at the same time. Note that some objects are dynamic, they can appear and disappear at any time or change in shape, become bigger or smaller or experience a major change in its original shape. This complicates the task of the data mining technique employed.

There are different models that can be explored depending on the how the data is collected. For instance the object can be well defined in the space and its locations are recorded in regular timestamps. Things can become more complicated when the object is not defined clearly and its location is not recorded at the same time interval. For the application of energy load forecasting, we have developed approaches based only on temporal analysis. In the case where the data collections exist for different regions, then it is worth looking at models taking into account the spatial dimensions. Currently, we have collections of data from different European countries. Our aim is to develop spatio-temporal models and evaluate both the accuracy and the complexity of such models.

To summarise, depending on the application at hand the knowledge discovery approach can be very complicated involving different inter-dependent steps to be dealt with. Until now, scientists and researchers have proposed concise solutions to specific problems. However, while these solutions work very well on the original targeted problems, they may, it is often the case, deliver poor performance on another problem. In section 6, we will describe a general framework for spatio-temporal data mining by trying to address their main challenges. We will leave very specific details of a given application to the user.

5. Visualisation for spatial data mining

Visualisation involves the use of visual/graphics techniques to represent information, data or knowledge. These techniques can be employed where complex datasets need to be explained or analyzed for developing and communicating conceptual information. The essential idea is that visual representations can help the user to get a better understanding of

content of the datasets, since the human visual system is more inclined to process visual rather than textual information. Thus, visualisation techniques may act as intelligence amplification tools for aiding and enhancing human intelligence, improving the perceptive, cognitive, and analytical abilities of people to allow them to solve complex tasks.
Nowadays, there exist very powerful computers able to quickly perform very complex and tedious tasks. Nevertheless, it is recognized that human performs better than computers in some areas, such as pattern recognition, evaluation, the overall sense of context that allows previously unrelated information to become related and useful, flexibility, and imagination.
Based on these considerations, visualisation is widely recognized as essential during data analysis and knowledge discovery for gaining an insight of data and underlying phenomena they represent, since it takes advantage of human abilities to perceive visual patterns and to interpret them (Andrienko et al., 2003), (Andrienko et al., 2005), (Johnston, 2001), (Kopanakis & Theodoulidis, 2003), (Costabile & Malerba, 2003). Moreover, it takes advantage of human ability to deal with non-completely defined problems (as often is the case for decision problems) thus overcoming an evident weakness of computers (Adrienko et al., 2007). As observed in (Walker, 1995):

"Programming a computer to "look for something interesting" in a database is a major undertaking, but given appropriate tools, it is a task for which humans are well equipped."

On the other hand, the powerful processing capabilities of computers are essential to deal with the huge amounts of data currently available. Thus, the idea of **Visual data mining** is to synergistically combine the computer processing capabilities with the unique and great human capabilities, to gain knowledge on the considered phenomena (Adrienko et al., 2007). In this context, a crucial role is played by the interaction techniques provided to the user for directly exploring the visual representation of data. Indeed, Visual Data Mining usually refers to methods, approaches and tools for the exploration of large datasets by allowing users to directly interact with visual representations of data and dynamically modify parameters to see how they affect the visualised data. Indeed, the usual approach follows the Information Seeking Mantra (Shneiderman, 1996) consisting of three steps: *Overview* first, *zoom and filter*, and then *details-on-demand*. Starting from an overview of the data the user identifies interesting patterns or subsets and focuses on one or more of them. Then, to analyse the identified patterns he/she requires to access details of the data exploiting a drill-down capability (Keim et al., 2003). For all the three steps of the process, effective visualisation techniques and interaction facilities have to be provided.
Visual Data Mining techniques, shifting the load from the user's cognitive system to the perceptual system, are able to enhance the effectiveness of the overall mining process, by supporting analytical reasoning, and have proven to be very valuable in many application domains.
In the context of **mining large spatial-temporal datasets** where geographical or physical space is involved, the visual exploratory approach is especially useful. Indeed:

- The heterogeneity of the space and the variety of properties and relationships in it cannot be adequately represented for fully automatic processing, thus there is the need to complement the computers capabilities with more sophisticated human capabilities.
- At the same time, an isomorphic visual representation, such as a map or an orthophoto, allows a human analyst or decision maker to perceive spatial relationships and patterns directly.

- Furthermore, a map or photo portraying coasts, rivers, relief, state boundaries, cities, roads, etc. exhibits not only the heterogeneity of the space but establishes the geographic context within which decisions can be made. The analyst or decision maker can grasp this information and relate it to his/her background knowledge about the properties of different parts of the space and take the variation of the properties into account.

However it is widely recognized that spatial visualisation features provided by existing geographical applications are not adequate for decision-support systems when used alone, but alternative solutions have to be defined. Existing tools (in particular, GIS, which are most commonly used as spatial decision aids) are often incapable to cope with the size and complexity of real-life problems, which forces the users to reduce the problems in order to adapt them to the capabilities of the tools (Adrienko et al. 2007). Moreover, as we have pointed out above, visualisation techniques for data exploration should not only include a static graphical view of the results produced by the mining algorithms, but also the possibility to dynamically obtain different spatial and temporal views as well as to interact in several ways with them. For example, the functionality of dynamically changing some of the involved parameters, and for that of quickly switching between different views for fast comparisons should be provided. This could allow the discovery of details and patterns that might remain hidden otherwise.

Some challenges can be identified for visual data mining of large spatial-temporal datasets:

- First of all, it is crucial to identify the most effective way to visualise the spatio-temporal multidimensional dataset taking into account the specific characteristics of the dataset, in order to communicate the useful and relevant information and to amplify the human capabilities.
- These visualisation methods and tools must be scalable with respect to the amount of data, dimensionality, number of data sources and heterogeneity of information, data quality and resolution, and characteristics of various displays and environments such as size, resolution, interaction possibilities, etc.
- It is important to provide effective visual interfaces for viewing and manipulating the geometrical and temporal attributes of the spatial-temporal data.
- Both the visualisation techniques and the interaction techniques should take into account that several persons are usually involved in the decision-making processes that data mining should support. These people have different role (administrators, politicians, data mining experts, domain experts, people affected by the decisions made) and very different background, thus requiring different needs. So, a user and task centered approach should be adopted to define appropriate visualisation and interaction techniques to effectively support each one of the identified actors and at the same time to allow them to fruitfully collaborate.

Visual data mining for spatio-temporal dataset is an interdisciplinary research area where techniques and expertise from information visualisation, visual perception, visual metaphors, diagrammatic reasoning, 3D computer graphics need to be suitably combined with the ones from data mining and geographic information systems.

6. Our approach

In this section we describe our work on the combination of visualisation and data mining techniques for spatio-temporal analysis and exploration.

As a case study and test bed we considered the Hurricane Isabel dataset (http://www.tpc.ncep.noaa.gov/2003isabel.shtml), which struck the US east coast in 2003

(National Hurricane Center 2003). This dataset is represented by a space of (500x500x100) with 25x106 real valued points in each of the 48 time-steps (approximately 62.5 GB). The main problem for analysing very large datasets such as this is that the hardware resources are not able to deal with the storage (memory) and processing (CPU) within the response time expected by the user to perform interactive queries.

To tackle this problem we have developed a 2-pass strategy (Di Martino et al 2006, Bertolotto et al., 2007). The goal of this type of strategy is to reduce the amount of memory used during the mining process as well as the processing time of a user query. Ideally, the data reduction or compression should not affect the knowledge contained in the data. The first task, then, in these strategies is to find the data points that are most similar according to their static (non spatial and temporal) parameters. This first phase is the key to the whole success of the compression, so that we do not lose any important information from the data that could have an adverse effect on the result obtained from mining the data at a later stage. The second task is to cluster these groups of closely related data points in a meaningful way to produce new "meta-data" sets that are more suitable and acceptable for data mining techniques to analyse and produce results (i.e. models, patterns, rules, etc.).

We have implemented two traditional clustering algorithms; DBSCAN and CURE (Kechadi et al. 2007). The first algorithm is more suitable for similarity measures that can be represented by a distance measure. Each cluster is represented by one data object. CURE can accept any similarity measure and the clusters can be represented by more than one representative. This is very important to represent clusters of different shapes. There are locations in the space that are highly similar; these are represented within each of the small location groups. It is important that no location group overlaps with another location group so that the integrity of the data is not affected.

In spatial and spatio-temporal data mining the effective analysis of results is crucial. Visualisation techniques are fundamental in the support of this task. We employed geo-visualisation in support of spatio-temporal data mining by developing two alternative interfaces, one based on the Google Earth application and one based on a Java 3D implementation. These tools have been developed to fit the complementary requirements posed by domain and mining experts, to allow the definition of a distributed, collaborative environment, and to deal with a dataset that could take advantage of a three-dimensional visualization in a geo-referenced space. The Google Earth-based tool renders in 3D the mining outcomes over a geo-referenced satellite image, enhanced by additional informative layers. The Java 3D-based tool provides more advanced user interaction with the mining results, by providing a set of features oriented to data-mining experts. These tools can be used in conjunction, as well as linked to any number of other instances, over an IP-based network, to create a collaborative and distributed environment.

In particular within our system, we developed an interacting visualisation functionality that allows not only to visualise the shape of clusters but also the shape of rules extracted by the mining algorithm. Therefore, a cluster or a rule produced at the mining layer is accessed directly and represented by its shape at the visualisation layer. This shape represents the region of space where the extracted rule holds (i.e., the set of locations in which the rule and hence all objects involved in it are well supported). While simply removing confusion and overload of visual information from the screen, it also help to highlight the structure of any pattern embedded in the data and to focus the user's attention only on the subset of the dataset involved in the rule being studied. This allows a more efficient and light visualization process, even when displaying millions of points.

The developed system has been successfully tested against the meteorological dataset, gathering positive results, since the system allowed us to detect both expected and

unexpected behaviours, as well as to find interesting relationships and specific patterns/characteristics about hurricane data (see Bertolotto et al., 2007 and Kechadi et al. 2007 for more details).

7. Conclusion

This chapter discusses available techniques and current research trends in the field of spatio-temporal data mining. An overview of the proposed approaches to deal with the spatial and the temporal aspects of data has been presented. Approaches that aim at taking into account both aspects were also surveyed.

Many challenges are still to be addressed. In particular, in the cartography and GIS community background/domain knowledge plays a significant role in the analysis of data. Therefore one of the biggest challenges is to integrate background geographic knowledge within the mining process. This is still an unexplored issue.

Currently, huge volumes of data are collected daily are often heterogeneous, geographically distributed and owned by different organisations. For example, an application by its nature is distributed such as an environmental application for which the data is collected in different locations and times using different instruments, and therefore these separate datasets may have different formats and features. So, traditional centralised data management and mining techniques are not adequate anymore. Distributed and high performance computing knowledge discovery techniques constitute a better alternative as they are scalable and can deal efficiently with data heterogeneity. So distributed data mining has become necessary for large and multi-scenario datasets requiring resources, which are heterogeneous and distributed. This constitutes an additional complexity to spatio-temporal data mining. We will look at this problem in our ADMIRE framework (Le-Khac 2006).

Visual techniques are essential for effective interpretation of mining results and as support to the mining process itself. We have discussed such techniques and presented our work in the area.

8. Acknowledgements

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

9. References

Andrienko N., Andrienko G., and Gatalsky P., Exploratory Spatio-Temporal Visualization: an Analytical Review. Journal of Visual Languages and Computing, special issue on Visual Data Mining. December 2003, v.14 (6), pp. 503-541.

Andrienko N., Andrienko G., Exploratory Analysis of Spatial and Temporal Data – A Systematic Approach, Springer, 2005.

Andrienko, G.L., Andrienko, N.V., Jankowski, P., Keim, D.A. , Kraak, M-J., MacEachren, A.M. Wrobel, S., Geovisual analytics for spatial decision support: Setting the research agenda. International Journal of Geographical Information Science 21(8): 839-857, 2007.

Bertolotto, M., Di Martino, S., Ferrucci, F., Kechadi, M-T., Visualization System for Collaborative Spatio-Temporal Data Mining, Journal of Geographical Information Science, Vol. 21, No. 7, July, 2007.

Camossi, E., Bertolotto, M., Kechadi, M-T., Mining Spatio-Temporal Data at Different Levels of Detail, Association Geographic Information Laboratories Europe, Girona, Spain, May 5-8, 2008.

Compieta, P., Di Martino, S., Bertolotto, M., Ferrucci, F., Kechadi, M-T., Exploratory Spatio-Temporal Data Mining and Visualization, Journal of Visual Languages and Computing , Vol. 18, No. 3, June 2007.

Cortes, C., Vapnik, K., Support Vector Networks, Machine Learning, 20(3): 273-297, 1995.

Costabile, M.F., Malerba, D. (Editors), Special Issue on Visual Data Mining, Journal of Visual Languages and Computing, Vol. 14, December 2003, 499-501.

Di Martino, S., Ferrucci, F., Bertolotto, M., and Kechadi, M-T., Towards a Flexible System for Exploratory Spatio-Temporal Data Mining and Visualization, Workshop on Visualization, Analytics & Spatial Decision Support (in GIScience'06), Münster, Germany, September 20-23, 2006.

Elman, J.L., Finding Structure in Time, Cognitive Science, Vol 14, No. 2, 1990, 179-211.

Ester, M., Kriegel, H.-P., Sander, J., Algorithms and applications for spatial data mining, in H. J. Miller and J. Han (eds.) Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, 160-187, 2001.

Ewens, W.J., Grant, G.R., Statistical methods in bioinformatics: An introduction (New York: Springer-Verlag), 2001.

Fitzgerald, T., Kechadi, M-T., Geiselbrechtinger, F., Application of fuzzy logic to online recognition of handwritten symbols, IEEE, 9th Int'l. Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan, October 26-29, 2004.

Gleeson, B. and Kechadi, M-T., Electric Load Forecasting Using Weather Data with a Kohonen Network and Data Mining Approach, The 26th International Symposium on Forecasting, Santander, Spain, June 11-14, 2006.

Gopal, S., Liu, W., Woodcock, X. Visualization based on fuzzy ARTMAP neural network for mining remotely sensed data, in H. J. Miller and J. Han (eds.) Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, 315-336, 2001.

Han, J., Kamber, M., Tung, A. K. H. (2001) "Spatial clustering methods in data mining: A survey," in H. J. Miller and J. Han (eds.) Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, 188-217.

Hinneburg, A., Keim, D. A., (1998) An efficient approach to clustering in large multimedia databases with noise, in Proceedings KDD, New York, 58-65.

Huang, B.Q., Rashid, T., Kechadi, M-T., A New Modified Network Based on the Elman Network, Int'l. Conference on Artificial Intelligence and Applications, Innsbruck, Austria, Feb. 16-18, 2004.

Huang, B.Q., Kechadi, M-T., A Recurrent Neural Network Recogniser for Online Recognition of Handwritten Symbols, The 7th Int'l. Conference on Enterprise Information Systems (ICEIS'05), Miami, Fl, USA, May 24-28, 2005.

Huang, B.Q., Rashid, T., Kechadi, M-T., Multi-Context Recurrent Neural Network Time Series Applications", International Journal of Computational Intelligence, Vol. 3, No. 3, February 2006.

Johnston W.L., Model visualization, in: Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, Los Altos, CA, 2001, pp. 223–227.

D.A. Keim, C.Panse, and M. Sips "Visual Data Mining of Large Spatial Data Sets", N. Bianchi-Berthouze (Ed.): DNIS 2003, LNCS 2822, pp. 201–215, 2003.

Kechadi, M-T., Bertolotto, M., Di Martino, S., Ferrucci, F., Scalable 2-Pass Data Mining Technique for Large Scale Spatio-Temporal Datasets, LNCS on Knowledge-Based Intelligent Information & Engineering Systems, 4693, 785-792, 2007.

Kohonen T., Self-Organising Maps, Springer Series in Information Sciences, Vol. 30, 1995.
Kopanakis I., Theodoulidis B., Visual data mining modeling techniques for the visualization of mining outcomes. Journal of Visual Languages and Computing. 14(6): 543-589, 2003.
Koperski K., Adhikary, J., Han, J., Spatial Data Mining: Progress and challenges, Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal, Canada, 55-70, 1996.
Koperski, K., Han, J., and Stefanovic N., An efficient two-step method for classification of spatial data, Proceedings of the Spatial Data Handling Conference, Vancouver, Canada, 1998.
Laxman, S., Sastry, P.S., Unnikrishnan K.P., Discovering Frequent Episodes and Learning Hidden Markov Models: A formal Connection, IEEE Trans. Knowledge data Eng., 17 1595-1517, 2005.
Le-Khac, N.A., Kechadi, M-T., ADMIRE Framework: Distributed Data Mining on Data-Grid Platforms, Int'l. Conference on Software and Data Technologies (ICSOFT'06), Setubal, Portugal, September 11-14, 2006.
Malerba, D., Esposito, F., Lanza, A., Lisi, F.A., Machine learning for information extraction from topographic maps, in H. J. Miller and J. Han (editors) Geographic Data Mining and Knowledge Discovery, Taylor and Francis, pp. 291–314, 2001.
Ng, R., Detecting outliers from large datasets, in H. J. Miller and J. Han (editors) Geographic Data Mining and Knowledge Discovery, Taylor and Francis, 218-235, 2001.
Peuquet, D.J. and Duan, N., An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. International journal of Geographic Information systems. 9:7-24, 1995.
Shneiderman, P., The eye have it: A task by data type taxonomy for information visualizations. In Visual Languages, 1996.
Picone, J., Continuous speech recognition using hidden Markov models, IEEE Signal processing magazine, 7:26-41, July 1990.
Qi, F. and Zhu, A.-X. Knowledge discovery from soil maps using inductive learning, International Journal of Geographical Information Science, 17, 771-795, 2003.
Rashid, T., Huang, B.Q., Kechadi, T-M., Auto-Regressive Recurrent Neural Network Approach for Electricity Load Forecasting, International Journal of Computational Intelligence, Vol. 3, No. 3, February 2006.
Roddick, J. F. and Lees, B., Paradigms for spatial and spatio-temporal data minig, in H. J. Miller and J. Han (eds.) Geographic Data Mining and Knowledge Discovery, London: Taylor and Francis, 33-49, 2001.
Shekhar, S., Huang, Y., Wu, W., Lu, C.T., Chawla, S., What's spatial about spatial data mining: three case studies, in R. Grossman, C. Kamath, V. Kumar, R. Namburu (editors), Data Mining for Scientific and Engineering Applications, Kluwer Academic Publishers, Dordrecht, 487-514, 2001.
Shekhar, S., Lu, C. T., Zhang, P., A unified approach to detecting spatial outliers, GeoInformatica, 7, 139-166, 2003.
Walker, G., Challenges in Information Visualisation, British Telecommunications Engineering Journal, Vol. 14, pp17-25, April 1995.
Yao, X., Research Issues in Spatio-Temporal Data Mining, workshop on Geospatial Visualization and Knowledge Discovery, Virginia, USA, Nov. 18-20, 2003.
Yuan, M., Use of knowledge acquisition to build wildfire representation in geographical information systems. International Journal of Geographic Information Science. 11:723-745, 1997.

Data Mining and Knowledge Discovery in Real Life Applications

Edited by Julio Ponce and Adem Karahoca

ISBN 978-3-902613-53-0
Hard cover, 436 pages
Publisher I-Tech Education and Publishing
Published online 01, January, 2009
Published in print edition January, 2009

This book presents four different ways of theoretical and practical advances and applications of data mining in different promising areas like Industrialist, Biological, and Social. Twenty six chapters cover different special topics with proposed novel ideas. Each chapter gives an overview of the subjects and some of the chapters have cases with offered data mining solutions. We hope that this book will be a useful aid in showing a right way for the students, researchers and practitioners in their studies.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

M-Tahar Kechadi, Michela Bertolotto, Filomena Ferrucci and Sergio Di Martino (2009). Mining Spatio-Temporal Datasets: Relevance, Challenges and Current Research Directions, Data Mining and Knowledge Discovery in Real Life Applications, Julio Ponce and Adem Karahoca (Ed.), ISBN: 978-3-902613-53-0, InTech, Available from: http://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/mining_spatio-temporal_datasets__relevance__challenges_and_current_research_directions

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen