A 2-Pass Data Mining Technique for Spatio-Temporal Datasets

M-T. Kechadi[1], M. Bertolotto[1], S. Di Martino[2], F. Ferrucci[2]
[1]University College Dublin, Belfield, Dublin 4, Ireland
Telephone: (+353) 1 7162913
Fax: (+353) 1 2697262
Email: tahar.kechadi@ucd.ie; michela.bertolotto@ucd.ie

[2]Università degli Studi di Salerno – DMI, Fisciano (SA), Italy
Telephone: (+39) 089 963374
Fax: (+39) 089 963303
Email: fferrucci@unisa.it; sdimartino@unisa.it

1. Introduction

Clustering is one of the fundamental techniques in data mining. It groups data objects based on characteristics of the objects and their relationships. It aims at maximizing the similarity within a group of objects and the dissimilarity between the groups in order to identify interesting structures in the underlying data. Several issues remain open in the clustering process. These include the optimised number of clusters, the validity of a given clustering, obtaining different shapes and sizes of clusters rather than forcing them into spherical shapes according to the distance measure functions, finding appropriate clustering structures in a given dataset.

Clustering algorithms can be divided into two main categories, namely partitioning and hierarchical methods. Partitioning methods divide the objects into *k* groups and iteratively exchange objects between the groups until no improvement can be made. Hierarchical clustering follows a bottom-up strategy that assumes the objects as the initial clusters and iteratively merges the closest pairs until the quality of the clusters does not further improve. BIRCH (Zhang et al. 1996), CURE (Guha et al. 1998), and C2P (Nanopoulos et al. 2001) are examples of hierarchical clustering methods. Furthermore, density-based approaches tend to cluster objects, which are close to each other and separate them from regions of low density. DBSCAN (Martin et al. 1996) and OPTICS (Ankerst et al. 1999) are the common algorithms of this class. Even though these algorithms are very popular they cannot be applied directly to very large spatio-temporal datasets. Indeed spatial and temporal constraints introduce a very high level of structure in the datasets that prevents most of the traditional data mining algorithms to discover models from such datasets. So far little work has been done in mining spatio-temporal data. Some modifications of the traditional methods have been proposed in (Vlachos et al. 2002) to cluster objects of similar trajectories. However, they are very computationally expensive and they cannot deal with distributed and heterogeneous aspects of the data.

In this paper, we describe a new method for mining very large spatio-temporal datasets. Because the raw dataset is too large for any algorithm to process, we reduce the size of that data by producing a smaller representation of the dataset so that it can be managed and mined interactively. Furthermore, we want to exploit the important aspect of spatio-temporal data (i.e., objects that are physically and temporally close tend to be "similar"). This data reduction is part of a 2-pass strategy, where the data objects are first

grouped according to their close similarity; these groups are then clustered by using different existing clustering techniques.

2. 2-Pass Strategy

The mining process for spatial data is more complex than for relational data in terms of both the mining efficiency and the complexity of possible patterns that can be extracted from spatial datasets (Roddick et al. 2001). The reason is that the attributes of neighbouring patterns may have significant influence on a pattern and should also be considered. Therefore, new techniques are required to efficiently and effectively mine spatial datasets.

The main problem for analysing very large datasets 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. For instance, Hurricane Isabel dataset (http://www.tpc.ncep.noaa.gov/2003isabel.shtml) is represented by a space of (500x500x100) with 25x106 data points. So the goal of this 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 this strategy is to find the data points that are most similar according to their static (non spatial and temporal) parameters. This part of the strategy is the key to the whole success of the process, 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 so that they are more suitable and acceptable for data mining techniques to analyse and produce results (i.e. models, patterns, rules, etc.).

For the first pass we deal with the raw data. This raw data can be represented by the following parameters: the number of time steps, the number of data points, and the number of static parameters. The algorithm for this first pass produces clusters of data points that are closely related. The goal here is to produce new data objects, where each object represents one cluster of raw data. It is important to note that only the data points that have a very high similarity between each other will be grouped together. As a result of this pass, the new dataset is much smaller than the original data. It contains more information about individual clusters. This will help the clustering performed during the second pass.

The second pass involves clustering the tightly grouped data objects to produce new clusters representing new knowledge and ready for evaluation and interpretation. We have implemented two traditional clustering algorithms; DBSCAN and CURE. The first one is more suitable for similarity measures that can be represented by a distance measure. Each cluster is represented by one data object, called medoid. The second one 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.

3. Preliminary Results

In our initial work on mining spatio-temporal datasets we have developed a 2-layer system architecture including a mining layer and a visualisation layer (Compieta et al. 2007, Bertolotto et al 2007). The mining layer implements a mining process along with the data preparation and interpretation steps. For instance, the data may need some cleaning and transformation according to possible constraints imposed by some tools, algorithms, or users. The interpretation step consists of using the selected models returned during the mining to effectively study the application behaviour. Usually, the model output requires some "post-processing" in order to exploit it. This step can benefit from data visualization, since interactivity and user expertise are very important in the final decision-making and data interpretation. Especially, in spatial data mining, the post-processing such as automatic and interactive visualisation of data is so important that some researchers considered it as part of modelling, and called it "Interactive Data Mining" (IDM) (Spenke et al. 1999). It combines both automatic and visual data mining, and received much attention from users as it offers higher degree of satisfaction and confidence (Keim et al. 2004).

Within our system, we developed an interacting visualisation tool 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.

We have implemented the 2-pass strategy presented in this paper within our system. We have conducted a small scale preliminary evaluation using the Hurricane Isabel datasets. Figure 1 displays an example of results obtained applying the DBSCAN technique within our 2-pass strategy. It outputs a very spherical cluster corresponding to the eye of the hurricane. It removes all the boundaries of the clusters.

We are currently completing the implementation of our system. The aim of generating and visualising clusters and rules is to provide some clues about direction or strength of the hurricane that can be useful for predicting the hurricane's behaviour. A more comprehensive evaluation is one of our goals for the near future. We also plan to apply and compare several different algorithms and improve the interactivity and flexibility of our visualisation tools.

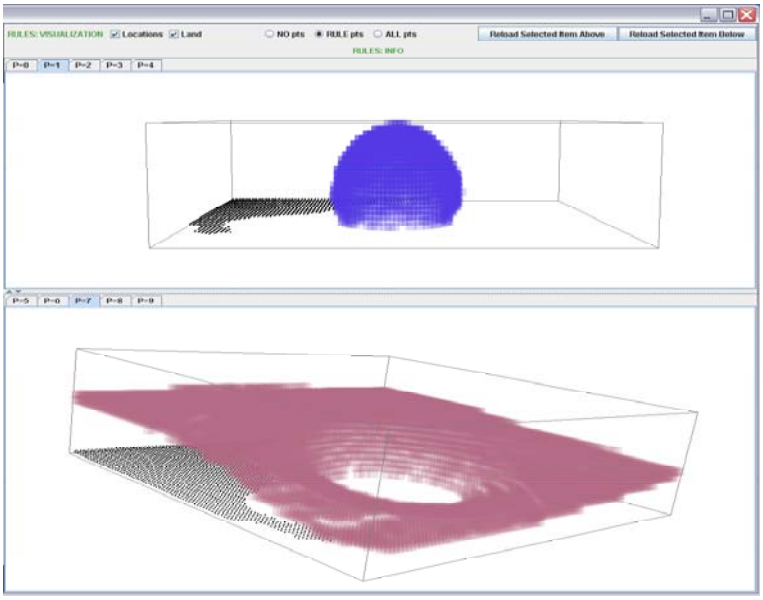

Figure 1. The eye of the hurricane in isolation (represented by one cluster)

4. References

Ankerst, M., Breunig, M., Kriegel, H.P., Sander, J, 1999, OPTICS: Ordering points to identify the clustering structure. In: Proc. of ACM SIGMOD, 49–60.

Bertolotto, M., Di Martino, S., Ferrucci, F., and Kechadi, T., 2007, A Visualization System for Collaborative Spatio-Temporal Data Mining, *International Journal of Geographical Information Science*, 21(7), Taylor & Francis.

Compieta, P., Di Martino, S., Bertolotto, M., Ferrucci, F., and Kechadi, T., 2007, Exploratory Spatio-Temporal Data Mining and Visualization, *Journal of Visual Languages and Computing*, Special Issue on Human-GIS Interaction, 18(3), Elsevier Science.

Guha, S., Rastogi, R., Shim, K., 1998, CURE: An efficient clustering algorithm for large databases, *Proceedings ACM SIGMOD'98*.

Keim D.A., Panse, C. , Sips, M., 2004, Visual Data Mining in Large Geospatial Poin Sets, *IEEE Computer Graphics and Applications*, 24(5), 36-44.

Martin, E., Kriegel, H.P., Sander, J., Xu, X., 1996, A density-based algorithm for discovering clusters in large spatial databases with noise, *Proceedings KDD'96*.

Nanopoulos, A., Theodoridis, Y., Manolopoulos, Y., 2001, C2P: Clustering based on closest pairs, *Proceedings VLDB'01*.

Roddick J.F., Lees, B.G. , 2001, Paradigms for spatial and spatio-temporal data mining, in H.G. Miller and J. Han (eds), Geographic Data Mining and Knowledge Discovery, London: Taylor & Francis.

Spenke M., C. Beilken, Visual, 1999, Interactive Data Mining with InfoZoom – The Financial Data Set, *Proceedings 3rd European Conference on Principles and Practice of Knowledge Discovery in databases, PKDD'99*, Sept. 15-18, Prague, Czech Republic.

Vlachos, M., Kollios, G., Gunopulos, D., 2002 Discovering similar multidimensional trajectories, *Proceedings ICDE'02*, 673–684.

Zhang, T., Ramakrishnan, R., Livny, M. , 1996, BIRCH: An efficient data clustering method for very large databases, *Proceedings ACM SIGMOD'96*.