Data Reduction in Very Large Spatio-Temporal Data Sets

Michael Whelan
School of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
e-mail: michael.whelan@ucd.ie

Nhien An Le Khac
School of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
e-mail: an.lekhac@ucd.ie

M-Tahar Kechadi
School of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
e-mail: tahar.kechadi@ucd.ie

***Abstract*— Today, huge amounts of data are being collected with spatial and temporal components from sources such as metrological, satellite imagery etc. Efficient visualisation as well as discovery of useful knowledge from these datasets is therefore very challenging and becoming a massive economic need. Data Mining has emerged as the technology to discover hidden knowledge from very large size of data. Furthermore, data mining techniques could be applied to decrease the large size of raw data by retrieving its useful knowledge as representatives. As a consequence, instead of dealing with a large size of raw data, we can use these representatives to visualise or to analyse without losing important information. This paper presents a data reduction technique based on clustering to help analyse very large spatio-temporal data. We also present and discuss preliminary results of this approach.**

Keywords-data mining; spatio-temporal datasets; clustering; data reduction

I. INTRODUCTION

Spatio-temporal data sets are often very large and difficult to analyse [8, 9, 10]. Since they are fundamental for decision support in many application contexts, recently a lot of interest has arisen towards data-mining techniques to filter out relevant subsets of very large data repositories as well as to help visualisation tools to effectively display results. Data mining techniques have been proven to be of significant value for spatio-temporal applications [12, 13]. It is a user-centric, interactive process, where data mining experts and domain experts work closely together to gain insight on a given problem. In particular, spatio-temporal data mining is an emerging research area, encompassing a set of exploratory, computational and interactive approaches for analysing very large spatial and spatio-temporal data sets. It is a combination of two widely researched areas, spatial data mining and temporal data mining. Spatial data mining is the discovery of interesting relationships and characteristics that may exist implicitly in spatial data sets. Temporal data mining concerns the analysis of events ordered by one or more dimensions of time. [12, 13] provide a very thorough set of references and surveys on the current state of the art research in all these areas. Several open issues have been identified ranging from the definition of mining techniques capable of dealing with spatio-temporal information to the development of effective methods for interpreting and presenting the final results.

Organisations are amassing very large repositories of customer, operations, scientific and other sorts of spatially and temporally related data of gigabytes or even terabytes size. Mining a database of even a few gigabytes is an arduous task for machine learning techniques and requires advanced parallel hardware and algorithms. Huge data sets create combinatorially explosive search spaces for data mining algorithms which may make the process of extracting useful knowledge infeasible owing to space and time constraints. The effectiveness of a technique for scaling data mining algorithms is measured in terms of the three factors, namely, time complexity, space complexity and quality of learning [8]. From the point of view of complexity analysis, for most scaling problems the limiting factor of the data set has been the number of examples and their dimension. A large number of examples introduces potential problems with both time and space complexity. For time complexity, the appropriate algorithmic question is what is the growth rate of the algorithm's run time as the number of examples and their dimensions increase? Also evaluating the effectiveness of a scaling technique becomes complicated if degradation in the quality of the learning is permitted. An approach for dealing with the intractable problem of learning from huge databases is to select a small subset of data for mining [9]. Databases often contain redundant data. It would be convenient if large databases could be replaced by a small subset of representative patterns so that the accuracy of estimates (e.g., of probability density, dependencies, class boundaries) obtained from such a reduced set should be comparable to that obtained using the entire data set.

Traditionally, the concept of Data Reduction has received several names, e.g. editing, condensing, filtering, thinning, etc, depending on the objective of the reduction task. Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same analytical results. There has been a lot of research into different techniques for the data reduction task which has lead to two different approaches depending on the overall objectives. The first one is to reduce the quantity of instances, while the second is to select a subset of features from the available ones. The later, known broadly as dimensionality reduction can be done in two ways, namely, feature selection and feature extraction. Feature selection refers to reducing the dimensionality of the space by

discarding redundant, dominated or least information carrying features. On the other hand, feature extraction methods utilise all the information contained in the data space to obtain a new transformed space, thereby mapping a higher dimensional pattern to a lower dimensional one. In this paper we will focus on the second approach to data reduction which deals with the reduction of the number of instances in the data set. Often called numerosity reduction or prototype selection, instance reduction algorithms are based on a distance calculation between instances in the data set. In such case selected instances, which are situated close to the centre of clusters of similar instances, serve as the reference instances. In this paper we focus on spatio-temporal clustering technique. Clustering is one of the fundamental techniques in data mining. It groups data objects based on information found in the data that describes the objects and their relationships. The goal is to optimise similarity within a group of objects and the dissimilarity between the groups in order to identify interesting structures in the underlying data. Clustering is used on spatio-temporal data to take advantage of the fact that, objects that are close together in space and/or in time can usually be grouped together. As a consequence, instead of dealing with a large size of raw data, we can use these cluster representatives to visualise or to analyse without losing important information.

The rest of the paper is organised as follows. In Section 2 we discuss related work. Section 3 describes in detail our data reduction technique based on clustering. Section 4 we evaluate the results of our data reduction technique as a pre-processing step on a very large spatio-temporal data set. In Sections 5 we discuss future work and concluded.

II. RELATED WORKS

A. Some Data Reduction Techniques

Sampling

The simplest approach for data reduction is to draw the desired number of random samples from the entire data set. Various random, deterministic and density biased sampling strategies exist in literature [1, 2]. However, naive sampling methods are not suitable for real world problems with noisy data, since the performance of the algorithms may change unpredictably and significantly. The random sampling approach effectively ignores all the information present in the samples not chosen for membership in the reduced subset. An advanced data reduction algorithm should include information from all samples in the reduction process [3, 4].

K-Nearest Neighbour

Some widely studied schemes for data reduction are built upon classification-based approaches, in general, and the k-nearest neighbour rule, in particular. The effectiveness of the reduced set is measured in terms of the classification accuracy. These methods attempt to derive a minimal consistent set, i.e., a minimal set which correctly classifies all the original samples. The very first development of this kind is the condensed nearest neighbour rule (CNN) [5]. Other algorithms in this category including the popular IB3, IB4 [6], reduced nearest neighbour and iterative condensation algorithms are summarised in [7].

Discretisation

Data discretisation techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data. Our current research focuses on improving the data reduction achieved by the discretisation technique [14]. In [14] the data reduction (HurricaneNarrower) consists of discretising numeric data into ordinal categories. The process starts by first being given a number of points (called split points or cut points) to split the entire attribute range, and then repeats this recursively on the resulting intervals.

B. Spatio-Temporal Data Mining System

Our strategy is to be incorporated in a system of exploratory spatio-temporal data mining [14,15], to improve its performance on very large spatio-temporal data sets. This system provides a data mining engine that can integrate different data mining algorithms and two complementary 3D visualisation tools. Within this system, there is a 2-layer architecture; a mining layer that provides newly developed techniques to efficiently support the data mining process, address the spatial and temporal dimensions of the data set, and a visualisation layer to visualise and interpret results. More details on the visualisation tools can be found in [14, 15].

In this paper we propose to improve two very important pre-processing utilities of this system called *Discretiser* and *Reducer*. The purpose of the Reducer utility is to deal with the very large size of the dataset that have to be analysed by the system. It works by scaling the datasets by a factor F; it simply runs through the whole dataset taking one average value for the F3 points inside each cube of edge F. Both these utilities have been found to be inefficient as a data reduction method which may lose a lot of important information contained in the raw data. So we propose a new data reduction method based on clustering to help with the mining of the very large spatio-temporal dataset. Because the raw data set is too large for any algorithm to process, the idea is to reduce the size of that data by producing a smaller representation of the data set, as opposed to compressing the data and then uncompressing it later for reuse. The reason is that we want to reduce and transform the data 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 firstly the data objects are grouped accordingly to their close similarity and secondly, these groups are then clustered by using different clustering techniques based on the clustering objectives. This strategy is used to design the clustering technique implemented in the first layer of our system.

III. CLUSTERING FOR COMPRESSION

Figure 1 shows an overview of our reduction method. The reason behind this method is to deal with the fact that the raw data is too large for any algorithm to process effectively, so the idea is to reduce the size of that data by producing a smaller representation of the data sets. The reduced data can then be analysed and produce useful information (i.e. models, patterns, rules, etc.) by applying other data mining techniques. 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"). From figure 1 the reduced data can then be analysed and produce useful information (i.e. models, patterns, rules, etc.) by applying other data mining techniques. We choose the data mining technique of clustering to assist in reducing the spatio-temporal datasets.

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 maximising the similarity within a group of objects and the dissimilarity between the groups in order to identify interesting structures in the underlying data. Some of the benefits of using clustering techniques to analyse spatio-temporal datasets included, a) the visualisation of clusters can help with understanding the structure of spatio-temporal data sets, b) the use of simplistic similarity measures to overcome the complexity of the datasets including the number of attributes, and c) the use of cluster representatives to help filter (reduce) datasets without losing important/interesting information.

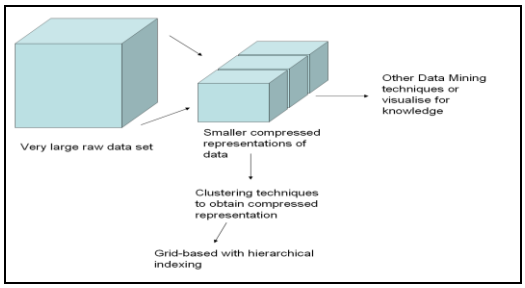


Figure 1. Overview of our mining strategy.

A. Clustering Algorithm

We have implemented the popular centre-based clustering method known as k-medoids. We have chosen a centre-based method rather than other clustering method such as density-based because of it simplicity. With centre-based clustering, you must specify a total number of passes. With each pass, the centres are adjusted to minimize the total distance between cluster centres and each record. The k-medoids algorithm chooses the closest data object to the centre of the cluster as the cluster representative. This is very important to our method as we can use this cluster medoid point's spatial and temporal attributes to visualise the clusters with their representatives (medoid points). This was the main advantage offered by k-medoids algorithm over other centre-based algorithms such as k-means, which would create new values for the cluster centre based on all the member of it cluster but would have no spatial or temporal attributes associated with it. So the goal here is to find data objects where each object represents one cluster of raw data (i.e. cluster representative).

We have chosen this simple clustering method for our initial run over the raw data as it is computationally simple and fast to complete. This part of our method 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. Figure 2 shows a high level view of the steps carried out by the algorithm for this first phase of this approach. 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 which then can be visualised by domain experts.

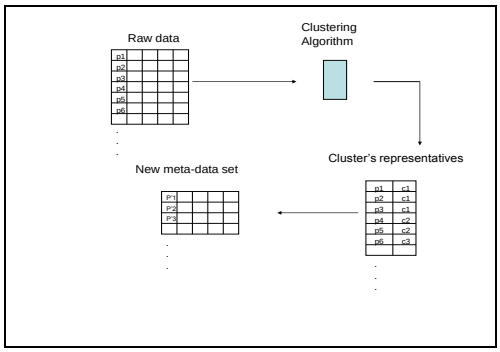


Figure 2. Step by step view of the first pass of the strategy.

IV. EVALUATION AND ANALYSIS

In this section, we study the feasibility of data reducing for spatio-temporal datasets by using data mining techniques described in Section III. The dataset is the Isabel hurricane data [11] produced by the US National Centre for Atmospheric Research (NCAR). It covers a period of 48 hours (time-steps). Each time-step contains several atmospheric variables. The grid resolution is 500×500×100. The total size of all files is more than 60GB (~ 1.25 GB for each time-step). The experimentation details and a discussion are given below.

A. Experimentation

The platform of our experimentation is a PC of 3.4Ghz Dual Core CPU, 3GB RAM using Java 1.6 on Linux kernel 2.6. Datasets of each time-step include 13 non-spatio attributes, so-called dimensions. In this evaluation, two dimensions are chosen for analysis: QVAPOR, the water vapour measured at each point of grid and P, weight of the atmosphere above a grid point. The range of QVAPOR value is [0…0.02368] and P is [-5471.85791…3225.42578]. We choose 3 time-steps 2, 10 and 18 to evaluate. We also filter the NULL value of testing datasets.

Figure 3, 4 and 5 show the (QVAPOR, P) in the grid coordinate at the selected time-steps. There are about 25 million data points for each time step. Figure 6, 7 and 8 show these (QVAPOR, P) dimensions after the reducing process by a centre-based clustering. We choose this technique because it is the most popular clustering technique in terms of the simplicity. The chosen clustering algorithm is K-Medoids. The number of clusters is 1000 and 2000. We only show the medoid point of each cluster as a representative.

B. Discussion

As shown in Figures 6, 7 and 8, 2000 representative points (or representative, in brief) could reflect the general shape of hurricane based on (QVAPOR, P) comparing to their whole original points (Figure 3, 4 and 5). Moreover, they can show different holes clearer then the original ones. These holes are normally very important to geography experts to study the interesting features of a hurricane. Furthermore, representatives also reflect the movement of hurricane as their whole origin points do.

However, these representatives can not reflect the data points on the left size of Figure 6, 7 and 8 (in the dash circles). The reason is that these points are border points of clusters not the medoid point in our approach. Besides, centre-based clustering techniques are not deal well with convex shape of datasets. We can improve this problem by either choosing initial points in dash circle area or apply density-based clustering techniques instead of centre-based one. However, running time as well as choosing efficient parameters for these techniques is also performance issues.

Moreover, there are only a small difference between the 2000 representatives and 1000 representatives (Figure 9, 10 and 11) in terms of general shape of the hurricane. So, in cases experts only needs to analyse the hurricane based on general shape and holes as well as running time is critical, we can use 1000 representatives instead of 2000 one.

These experiments show that simple data mining techniques can be applied to reduce the large size of spatio-temporal datasets in keeping of their important information used by experts.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we study the feasibility of using data mining techniques in reducing the large size of spatio-temporal datasets. As there are many reducing techniques presented in the literature such as discretisation, etc., most of them are concerned with reducing the dataset size without paying attention to their geographic properties. Besides, analysing spatio-temporal datasets is difficult by its nature. Recently, data mining has been raised as a technique to retrieve hidden knowledge in the large size of datasets. So, we propose to apply a clustering technique to reduce the large size without loosing important information. We apply K-Medoids clustering on different time-steps. The experimental results show that knowledge extracted from mining process can be used as efficient representatives of huge datasets.

We are testing with a hybrid approach where density-based and centre-based clustering's are used to increase the performance in terms of running time and representative positions. Besides, parallel and distributed clustering techniques are also being studied to analyse all non-spatio attributes in order to prove the robustness of our approach.

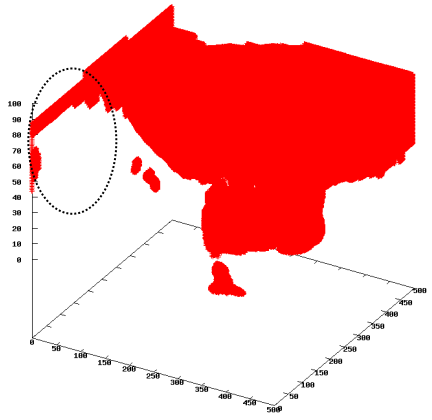


Figure 3. (QVAPOR, P) at Time-step 2.

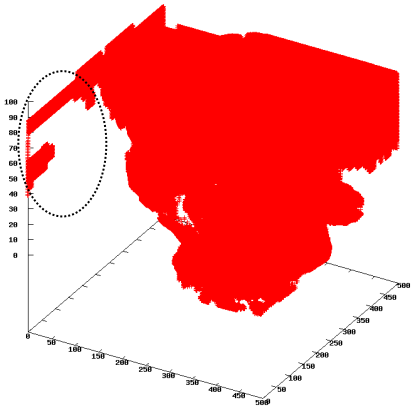
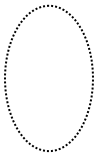


Figure 4. (QVAPOR, P) at Time-step 10.



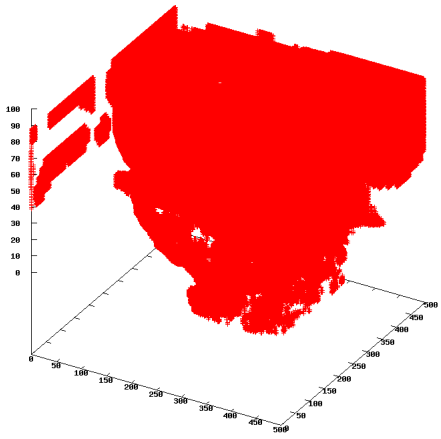


Figure 5. (QVAPOR, P) at Time-step 18.

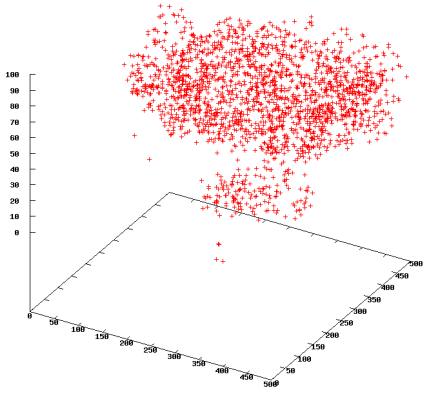


Figure 6. 2000 representatives of (QVAPOR, P) at Time-step 2.

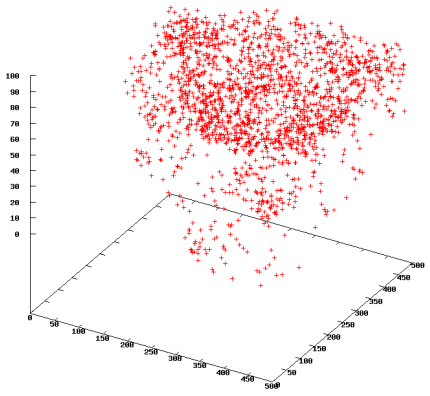


Figure 7. 2000 representatives of (QVAPOR, P) at Time-step 10.

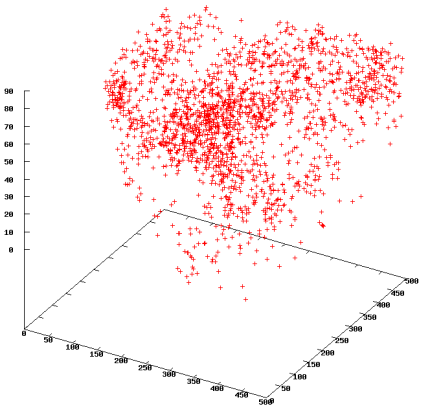


Figure 8. 2000 representatives of (QVAPOR, P) at Time-step 18.

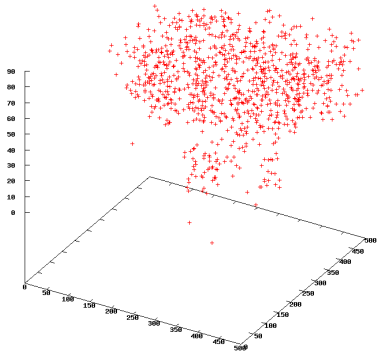


Figure 9. 1000 representatives of (QVAPOR, P) at Time-step 2.

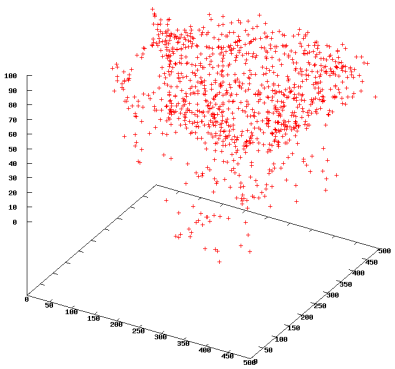


Figure 10. 1000 representatives of (QVAPOR, P) at Time-step 10.

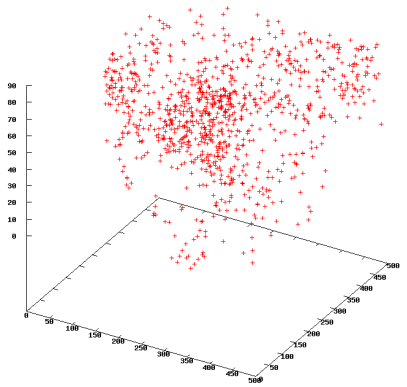


Figure 11. 1000 representatives of (QVAPOR,P) at Time-step 18.

REFERENCES

[1] Kivinen, J. and Mannila, H. 1994. The power of sampling in knowledge discovery. In Proc. ACM SIGACT-SIGMOD-SIGART (Minneapolis, Minnesota, United States, May 24 - 27, 1994). PODS '94. ACM, New York, NY, 77-85.

[2] Liu, H. and Motoda, H. 2002. On Issues of Instance Selection. Data Min. Knowl. Discov. 6, 2 (Apr. 2002), 115-130.

[3] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning, In Proc. of ICML-94, 11th International Conference on Machine LearningNew Brunswick, US: Morgan Kaufmann Publishers, San Francisco, US (1994) , p. 148--156.

[4] Cohn, D., Atlas, L., and Ladner, R. 1994. Improving Generalization with Active Learning. Mach. Learn. 15, 2 (May. 1994), 201-221.

[5] Angiulli, F. 2005. Fast condensed nearest neighbor rule. In Proc. 22nd international Conference on Machine Learning (Bonn, Germany, August 07 - 11, 2005). ICML '05, vol. 119. ACM, New York, NY, 25-32.

[6] Aha, D. W. 1992. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. Int. J. Man-Mach. Stud. 36, 2 (Feb. 1992), 267-287.

[7] Wilson, D.R., Martinez, T.R.: Reduction Techniques for Instance-based Learning Algorithm. Machine Learning 33-3, 257–286 (2000).

[8] Dunham M. H., "Data Mining: Introductory and Advanced Topics", Prentice Hall, 2003.

[9] Tan, P-N., Steinbach, M. and Kumar, V., "Introduction to Data Mining", Addison Wesley, 2006.

[10] Ye N (ed) The Handbook of Data Mining. Lawrence Erlbaum Associates Publishers, Mahwah, New Jersey, USA 2003.

[11] National Hurricane Center, "Tropical Cyclone Report: Hurricane Isabel", http://www.tpc.ncep.noaa.gov/2003isabel.shtml, 2003.

[12] Roddick, J. F., Hornsby, K., and Spiliopoulou, M. 2001. An Updated Bibliography of Temporal, Spatial, and Spatio-temporal Data Mining Research. In Proc. 1st international Workshop on Temporal, Spatial, and Spatio-Temporal Data Mining-Revised Papers J. F. Roddick and K. Hornsby, Eds. Lecture Notes In Computer Science, vol. 2007. Springer-Verlag, London, 147-164.

[13] Roddick J.F., Lees B.G., "Paradigms for Spatial and Spatio-Temporal Data Mining", in H. J. Miller and J. Han: Geographic Data Mining and Knowledge Discovery (London: Taylor and Francis), 2001.

[14] Compieta, P., Di Martino, S., Bertolotto, M., Ferrucci, F., and Kechadi, T. 2007. Exploratory spatio-temporal data mining and visualization. J. Vis. Lang. Comput. 18, 3 (Jun. 2007), 255-279.

[15] Bertolotto, M., Di Martino, S., Ferrucci, F., and Kechadi, T., " A Visualisation System for Collaborative Spatio-Temporal Data Mining", International Journal of Geographical Information Science, Vol. 21, No. 7, July, 2007.