National Workshop-Cum-Conference on Recent Trends in Mathematics and Computing (RTMC) 2011
Proceedings published in International Journal of Computer Applications® (IJCA)

An Improved Clustering Approach on Time Series Data Set

Pallavi [1], Sunila Godara [2]

Student ,Department of computer science & engg. ,GJU S&T, Hisar[1]
Assistant professor, Department of computer science & engg ., GJU S&T, Hisar[2]

ABSTRACT

In clustering, objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. This paper proposes use of BIRCH hierarchical clustering method on large amount of numerical data by integration of hierarchical clustering and other clustering methods such as iterative partitioning methods such as k-means and k-medoids and their comparison. Clustering feature and clustering feature tree (CF tree) will be used to summarize cluster representations. With this clustering method we can achieve good speed and scalability.

1. INTRODUCTION

Knowledge discovery process consists of an iterative sequence of steps such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Data mining functionalities are characterization and discrimination, mining frequent patterns, association, correlation, classification and prediction, cluster analysis, outlier analysis and evolution analysis [1].

Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another and very dissimilar to object in other clusters. Dissimilarity is based on the attributes values describing the objects. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. Firstly the set of data is portioned into groups based on data similarity (e g Using clustering) and the then assign labels to the relatively small number of groups.

Several clustering techniques are there: partitioning methods, hierarchical methods, density based methods, grid based methods, model based methods, methods for high dimensional data and constraint based clustering.

Clustering is also called data segmentation because clustering partitions large data sets into groups according to their similarity. Clustering can be used for outlier detection where outliers may be more interesting then common cases e g Credit card fraud detection, monitoring of criminal activities in electronic commerce. Clustering is a pre-processing step for other algorithms such as characterization, attribute subset selection and classification, which would then operate on the detected clusters and the selected attributes or features.

Contributing area of research include data mining, statistics, machine learning, special database technology, biology and marketing. Clustering is an unsupervised learning. Unlike classification, it does not rely on predefined classes and class labels training examples. Hence we say clustering is learning by observation rather than learning by examples.

Typical requirements of clustering in data mining are Scalability, ability to deal with different types of attributes, Discovery of clusters with different shapes, Minimal requirement for domain knowledge to determine input parameters, Ability to deal with noisy data, Incremental Clustering and insensitivity to the order of input records, High dimensionality, Constraint based Clustering and Interpretability and usability.

2. HIERARCHICAL CLUSTERING

Hierarchical clustering methods works by grouping data objects into a tree of clusters. Hierarchical clustering methods can be further classified as agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom up (merging) or top down (splitting) fashion.

BIRCH: Balanced iterative reducing and clustering using hierarchies. BIRCH is designed for clustering a large amount of numerical data by integration of hierarchical clustering (at the initial micro clustering stage) and other clustering methods such as iterative partitioning (at the later macro clustering stage). It overcomes the two difficulties of agglomerative clustering methods: scalability and the inability to undo what was done in the previous step.

BIRCH introduces two concepts, clustering feature and clustering feature tree (CF tree) which are used to summarize cluster representation. These structures help the clustering method achieve good speed and scalability in large databases and also make it effective for incremental and dynamic clustering of incoming objects.

3. TIME SERIES DATABASE

A time-series database consists of sequences of values or events obtained over repeated measurements of time. The values are typically measured at equal time intervals (e.g., hourly, daily, weekly). Time-series databases are popular in many applications, such as: Stock Market Analysis, Economic and Sales Forecasting, Budgetary Analysis, Utility Studies, Inventory Studies, Yield Projections, Workload Projections, Process and Quality control, Observation of Natural Phenomena (such as atmosphere, temperature, wind, earthquake), Scientific and Engineering Experiments, Medical Treatments.

A time-series database is also a sequence database. Sequence database is any database that consists of sequences of ordered events, with or without concrete notions of time. For example, Web page traversal sequences and customer shopping transaction sequences are sequence data, but they may not be time-series data.

Our major issues in mining Time Series are How can we find correlation relationships within time-series data? How can we analyse such huge numbers of time series to find similar or regular patterns, trends, bursts (such as sudden sharp changes), and outliers, with fast or even on-line real-time response? This has become an increasingly important and challenging problem.

The proposed paper explores the use of BIRCH Hierarchical Clustering method with coordination of iterative partitioning methods such as K-Means and K-Medoid and then we would compare results of our method with the existing techniques.

4. PREVIOUS WORK AND ANALYSIS

Xiaozhe Wang et al., in 2006 provide a method for clustering of time series based on their structural characteristics, rather it clusters based on global features extracted from the time series. Global measures describing the time series are obtained by applying statistical operations that best capture the underlying characteristics: trend, seasonality, periodicity, serial correlation, skewness, kurtosis, chaos, nonlinearity, and self-similarity. Since the method clusters using extracted global measures, it reduces the dimensionality of the time series and is much less sensitive to missing or noisy data. A search mechanism is provided to find the best selection from the feature set that should be used as the clustering inputs [2].

Li Wei et al. in 2005 stated a practical tool for visualizing and data mining medical time series and concluded that increasing interest in time series data mining has had surprisingly little impact on real world medical applications. Practitioners who work with time series on a daily basis rarely take advantage of the wealth of tools that the data mining community has made available. This approach extracts features from a time series of arbitrary length and uses information about the relative frequency of these features to colour a bitmap in a principled way. By visualizing the similarities and differences within a collection of bitmaps, a user can quickly discover clusters, anomalies, and other regularities within the data collection [3].

An Online Algorithm for Segmenting Time series was carried out by Eamonn Keogh et al., in 2001. This was the first extensive review and empirical comparison of time series segmentation algorithms from a data mining perspective. This showed the most popular approach, Sliding Windows, generally produces very poor results, and that while the second most popular approach, Top-Down, can produce reasonable results, it does not scale well. In contrast, the least well known, Bottom-Up, approach produces excellent results and scales linearly with the size of the dataset. In addition, this introduced SWAB, a new online algorithm, which scales linearly with the size of the dataset, requires only constant space and produces high quality approximations of the data [4].

A Model Based Clustering For Time Series With Irregular Interval was proposed by Xiao-Tao Zhang et al., in 2004. This focussed Clustering problems are central to many knowledge discovery and data mining tasks. However, most existing clustering methods can only work with fixed-interval representations of data patterns, ignoring the variance of time axis. This studied the clustering of data patterns that are sample in irregular interval. A model-based approach using cepstnun distance metric and Autoregressive Conditional Duration (ACD) model has proposed. Experimental results on real datasets showed that this method is generally effective in clustering irregular space time series, and conclusion inferred from experimental results agrees with the market microstructure theories. [5]

Hui Ding et al., in 2008 experimentally compared the representations and distance measures of querying and mining of Time Series Data. This conducted an extensive set of time series experiments re-implementing 8 different representation methods and 9 similarity measures and their variants, and testing their effectiveness on 38 time series data sets from a wide variety of application domains. They gave an overview of these different techniques and present their comparative experimental findings regarding their effectiveness. Their experiments have provided both a unified validation of some of the existing achievements, and in some cases, suggested that certain claims in the literature may be unduly optimistic [6].

Ehsan Hajizadeh et al.,in 2010 provided an overview of application of data mining techniques such as decision tree, neural network, association rules, factor analysis and etc in stock markets. Also, this reveals progressive applications in addition to existing gap and less considered area and determines the future works for researchers. This stated problems of data mining in finance (stock market) and specific requirements for data mining methods including in making interpretations, incorporating relations and probabilistic learning. The data mining techniques outlined here advances pattern discovery methods that deals with complex numeric and non-numeric data, involving structured objects, text and data in a variety of discrete and continuous scales (nominal, order, absolute and so on). Also, this show benefits of using such techniques for stock market forecast [7].

Jiangjiao Duan et al., in 2005 introduced that Model-based clustering is one of the most important ways for time series data mining. However, the process of clustering may encounter several problems. Here a novel clustering algorithm of time-series which incorporates recursive Hidden Markov Model (HMM) training was proposed. It contributed the following aspects: 1) It recursively train models and use this model information in the process agglomerative hierarchical clustering. 2) It built HMM of time series clusters to describe clusters. To evaluate the effectiveness of the algorithm, several experiments had conducted on both synthetic data and real world data. The result shows that this approach can achieve better performance in correctness rate than the traditional HMM-based clustering algorithm [8].

Information Mining Over Heterogeneous and High-Dimensional Time-Series Data in Clinical Trials Databases was carried out by Fatih Altiparmak et al., in 2006. They gave a novel approach for information mining that involves two major steps: applying a data mining algorithm over homogeneous subsets of data, and identifying common or distinct patterns over the information gathered in the first step. This approach implemented specifically for heterogeneous and high dimensional time series clinical trials data. Using this framework, this propose a new way of utilizing frequent item set mining, as well as clustering and declustering techniques with novel distance metrics for measuring similarity between time series data. By clustering the data, it find groups of analytes(substances in blood) that are most strongly correlated. Most of these relationships already known are verified by the clinical panels, and, in addition, they identify novel groups that need further biomedical analysis. A slight modification to this algorithm results an effective declustering of high dimensional time series data, which is then used for "feature selection." Using industry-sponsored clinical trials data sets, they are able to identify a small set of analytes that effectively models the state of normal health [9]

A Qualitative Feature Extraction Method for Time Series Analysis studies by Jinfei Xie & Wei-Yong Yan, in 2006. They told that Time series feature extraction is a way to reveal the most important characteristics of a (or a set of) time series. It is an effective pre-processing step for many time series mining tasks such as clustering and indexing. This proposed a new qualitative feature extraction method. The method differs from most available methods in that it mainly focuses on the shape, instead of the actual values, of any time series. In this method, a set of shape oriented patterns is defined and the feature of a data sequence is referred to as the combination of these patterns. A procedure for identifying patterns in a given sequence is developed. Experiments on real stock price data are performed to evaluate the performance of this method used for clustering and similarity search [10].

Cluster Time Series Based on Partial Information addresses the problem using a portion of information in clustering time series. It describe model for retrieving and representing the partial Information in time series data. It evaluates this approach through comparing the results with a standard classification. The result shows this approach could outperform the previous clustering method that is based on the whole information of time series [11].

5. PROPOSED ALGORITHM

Hierarchical clustering methods works by grouping data objects into a tree of clusters. Hierarchical clustering methods can be further classified as agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom up (merging) or top down (splitting) fashion. BIRCH: Balanced iterative reducing and clustering using hierarchies. BIRCH is designed for clustering a large amount of numerical data by integration of hierarchical clustering (at the initial micro clustering stage) and other clustering methods such as iterative partitioning (at the later macro clustering stage). It overcomes the two difficulties of agglomerative clustering methods: scalability and the inability to undo what was done in the previous step. BIRCH introduces two concepts, clustering feature and clustering feature tree (CF tree) which are used to summarize cluster representation. These structures help the clustering method achieve good speed and scalability in large databases and also make it effective for incremental and dynamic clustering of incoming objects. Given n d-dimensional data objects or points in a cluster, we can define the centroid x0, radius R, and diameter D of the cluster as follows:

$$x_0 = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$R = \sqrt{\frac{\sum_{i=1}^{n} (x_i - x_0)^2}{n}}$$

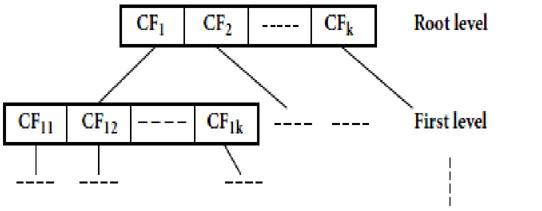
$$D = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2}{n(n-1)}}$$

Where R is the average distance from member objects to the centroid, and D is the average pair wise distance within a cluster. Both R and D reflect the tightness of the cluster around the centroid. A clustering feature (CF) is a three-dimensional vector summarizing information about clusters of objects. Given n d-dimensional objects or points in a cluster, fig, then the CF of the cluster is defined as

CF = <n, LS, SS>

Where n is the number of points in the cluster, LS is the linear sum of the n points and SS is the square sum of the data points .

Phase 1: BIRCH scans the database to build an initial in-memory CF tree, which can be viewed as a multilevel compression of the data that tries to preserve the inherent clustering structure of the data.



A CF tree structure.

Phase 2: BIRCH applies a (selected) clustering algorithm to cluster the leaf nodes of the CF tree, which removes sparse clusters as outliers and groups dense clusters into larger ones [1].

6. CONCLUSION

The hierarchical clustering method, though simple, often encounters difficulties regarding the selection of merge or split points. Such a decision is critical because once a group of objects is merged or split, the process at the next step will operate on the newly generated clusters. It will neither undo what was done previously nor perform object swapping between clusters. Thus merge or split decisions, if not well chosen at some step, may lead to low-quality clusters. Moreover, the method does not scale well, because each decision to merge or split requires the examination and evaluatation of a good number of objects or clusters. One promising direction for improving the clustering quality of hierarchical methods is to integrate hierarchical clustering with other clustering techniques, resulting in multiple-phase clustering, therefore BIRCH was introduced. BIRCH, begins by partitioning objects hierarchically using tree structures, where the leaf or low-level nonleaf nodes can be viewed as “microclusters” depending on the scale of resolution. It then applies other clustering algorithms to perform macroclustering on the microclusters

This paper proposes use of BIRCH hierarchical clustering method on large amount of numerical data by integration of hierarchical clustering and other clustering methods such as iterative partitioning methods such as k-means and k-medoids and their comparison. Clustering feature and clustering feature tree (CF tree) will be used to summarize cluster representations. With this clustering method we can achieve good speed and scalability.

7. REFERENCES

[1] Han J. and Kamber M.: "Data Mining: Concepts and Techniques," *Morgan Kaufmann Publishers,* San Francisco, 2000.

[2] Xiaozhe Wang, Kate Smith and Rob Hyndman: " Characteristic-Based Clustering for Time Series Data", *Data Mining and Knowledge Discovery , Springer Science + Business Media, LLC. Manufactured in the United States, 335–364, 2006.*

[3] Li Wei, Nitin Kumar,Venkata Lolla and Helga Van Herle:"A practical tool for visualizing and data mining medical time series", *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)* 106-125, 2005.

[4] Eamonn Keogh, Selina Chu,David Hart and Michael Pazzani:" An online algorithm for segmenting time series", *0-7695-1 119-8/01 IEEE,* 2001.

[5] Xiao-Tao Zhang, Wei Zhang and Xiong Xiong:" A model based clustering for time-series with irregular interval", *Proceedings of the Third International Conference on Machine Learhg and Cybernetics, Shanghai,* 26-29, **August** 2004.

[6] Hui Ding, Goce Trajcevski and Eamonn Keogh:" Querying and mining of time series data:Experimental comparison of representations and distance measures" , *PVLDB '08, August 23-28, 2008, Auckland, New Zealand,* 2008.

[7] Ehsan Hajizadeh, Hamed Davari Ardakani and Jamal Shahrabi:"Appilication of data mining techniques in stock market", *Journal of Economics and International Finance Vol. 2(7),* pp. 109-118, July 2010.

[8]Jiangjiao Duan, WeiWang , Bing Liu and Baile Shi:" Incorporating with recursive model training in time series clustering", *Proceedings of the 2005 The Fifth International Conference on Computer and Information Technology (CIT'05),* IEEE2005.

[9] Fatih Altiparmak, Hakan Ferhatosmanoglu, Selnur Erdal, and Donald C. TrostFaith Altipar:

" Information Mining Over Heterogeneous and High-Dimensional Time-Series Data in Clinical Trials Databases", *IEEE Transactions On Information Technology In BioMedicine,* VOL. 10, 215-239, APRIL 2006.

[10] Jinfei Xie, Wei-Yong Yan:" A Qualitative Feature Extraction Method for Time Series Analysis", *Proceedings of the 25th Chinese Control Conference 7–11 August, 2006,* Harbin, Heilongjiang, 2006.

[11] Xiaoming lin, Yuchang Lu, Chunyi Shi:" Cluster Time Series Based on Partial Information", *IEEE SMC TPUl,* p. no. 254-262, year 2002.