International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume III, Issue X, October- 2016

Temporal Data Clustering Framework with Optimal Distance Measures

Mr. G. Sakthivel, MCA., M.Phil, Research Scholar,
Mrs. J.S. Subhashini, MCA., M.Phil., Assistant Professor,
Department of Computer Science,
SSM College of Arts & Science, Komarapaliyam, Tamilnadu, India

Abstract

Data clustering methods are adapted to partition the similar data values. Distance measures are used to estimate the transaction similarity levels. Temporal databases are build with time series information. Time series data comparison is a complex task. Different features are considered in the time series data comparison process. Shift, seasonality, trend, correlation, skewness and auto correlation features are analyzed for the time series data values. Multi class classifiers are applied to select the distance measures for the time series data values.

The time series data clustering process is composed with Euclidean Distance (ED), Dynamic Time Warping (DTW), Edit Distance for Real Sequences (EDR), TQuest and Fourier Coefficient based Similarity Measure. The distance measure selection process is carried out with the Classifier Chain algorithm (CC) and Random-k-labelsets classifier (RkL). The partitional and hierarchical data clustering operations are carried out on the time series data values. The K-medoids clustering algorithm and agglomerative hierarchical algorithm are applied for the clustering process.

The uncertain data handling process is integrated with the time series data clustering process. Data dimensionality is balanced using the uncertain data handling process. The Inter Cluster Distance (ICD) based centroid estimation is adapted in the system. The Centroid Optimized K-Means (COKM) clustering scheme is build with Euclidean Distance (ED) measure. The Distance and Centroid Optimized K-Means (DCOKM) clustering scheme is constructed with Edit Distance for Real Sequences (EDR) measures. The classifier performs the distance measure selection for the clustering process.

1. Introduction

Data in the form of time series pervades a large number of scientific do mains. Observations that unfold over time usually represent valuable information subject to analysis, classification, in dexing, prediction, or interpretation. Real-world examples include financial data, medical data, computer data, or motion data. Even object shapes or handwriting can be effectively transformed into time series, facilitating their analysis and retrieval.

A core issue when dealing with time series is determining their pair wise similarity, i.e., the degree to which a given time series resembles an other. In fact, a time series similarity measure is central to many mining, retrieval, clustering, and classification tasks. Furthermore, there is evidence that simple approaches to such tasks exploiting generic time series similarity measures usually outperform more elaborate, some times specifically-targeted strategies. This is the case, for instance, with time series classification, where a one-nearest neighbor approach using a well-known time series similarity measure was found to outperform an causative list of alternatives, including decision trees, multi scale histograms, multi-layer perception neural networks, order logic rules with boosting, or multiple classifier systems.

Deriving a measure that correctly reects time series similarities is not straightforward. Apart from dealing with high dimensionality, the calculation of such measures needs to be fast and efficient. Indeed, with better information gathering tools, the size of time series data sets may continue to increase in the future. Moreover, there is the need for, generic/multi-purpose similarity measures, so that they can be readily applied to any data set, whether this application is the final goal or just an initial approach to a given task. This last aspect highlights another desirable quality for time series similarity measures: their robustness to different types of data.

2. Related Work

Simultaneous row and column clustering for identifying block structures from matrix data has been initially studied [12]. Recent surge of interests in co-clustering is motivated by biological applications, which aim at identifying subset of genes co-expressed in a subset of samples from microarray gene expression data. Co-clustering has also been applied in many other applications, including simultaneous clustering of words and documents, authors and conference, etc. Early work on co-clustering focuses on defining an error measure and then identifying blocks that minimize this measure using heuristic search algorithms. These early work has recently been reformulated using matrix and optimization techniques. Following the spectral clustering formalism, it has been shown recently that co-clustering is closely related to the SVD of the data

matrix. Co-clustering is formulated as a bipartite graph cut problem, and the data are projected onto the left and right singular vector spaces before they are concatenated and clustered to identify row and column co-clusters. It is shown in [11] that sparsityinducing regularization can be employed to compute sparse singular vectors, which in turn can be used to form co-clusters. In [6], a framework for simultaneous co-clustering and predictive learning is proposed.

This work is also related to recent studies on mining from time-evolving data. Chakrabarti et al. first proposed the concept of evolutionary clustering and extended the K-means and the hierarchical clustering algorithms for uncovering smooth patterns from time-evolving data matrices. The spectral clustering formalism is systematically extended to the evolutionary setting by incorporating a temporal cost into the objective function, leading to a suite of formulations for evolutionary spectral clustering. The nonnegative matrix factorization is employed for soft clustering, and a temporal cost is included for mining from timeevolving data. Evolutionary nonnegative matrix factorization is studied in [4], and the idea of adaptively estimating the smoothness parameter is proposed in [7]. The broad area of evolutionary network analysis is reviewed in [1].

The fused Lasso penalty was originally proposed for encouraging smoothness over related coefficients in regression problems. This type of penalty is very attractive and has been applied for encouraging smoothness over spatial and temporal smoothness in many applications, including biological data analysis and social studies. A critical challenge in employing the fused Lasso formalism is that this class of penalty is non-smooth and non-separable and thus is very challenging to optimize. A modified coordinate descent algorithm is developed to solve the fused Lasso formulation. This algorithm is not guaranteed to give the exact solution. In [9], a path algorithm is proposed to solve the fused Lasso signal approximator. Instead of solving the original primal problem, Liu et al. developed a dual formulation for the fused Lasso signal approximator and devised a gradient descent algorithm for computing the dual solution [2]. Similar formulations and algorithms have been studied in the compressive sensing literature [8], [10].

The problem of feature selection in clustering has been studied in [3], [5].. These studies mostly focus on clustering static data matrices. In the literature, the evolutionary clustering paradigm is related, but different from, the currently studied evolutionary feature selection formalism. Specifically, the smoothness constraints are imposed on the sample dimension in evolutionary clustering, while similar constraints are imposed on the feature dimension in evolutionary feature selection. Consequently, the clustering results are expected to evolve smoothly in evolutionary clustering, while the selected features are shared across time points in evolutionary feature selection.

3. Clustering Time Series Databases

In recent years, the increase in data collecting devices has enabled access to a large amount of temporal data. This has generated a new type of databases where each instance consists of an entire time series. The main characteristics of this type of data are its high dimensionality, its dynamism, its auto-correlation and its noisy nature, all of which complicate the analysis to a large extent. In view of this, many researchers have focused on finding new methods or on adapting the data mining algorithms to obtain useful information from these databases. For example, tasks such as clustering have been successfully adapted to time series data in many application domains.

The clustering of time series databases typically requires the definition of a distance measure which will estimate the level of similarity or dissimilarity between time series. Euclidean distance and other common measures used for nontemporal data are not always the most suitable methods to evaluate the similarity between time series, because they are unable to deal with noise and misalignments in the series. The scientific community has used a vast portfolio of measures for this specific type of data.

Experiments suggest that not all of these distance measures are appropriate for all time series databases. This is probably due to the specific characteristics of each database, which make some distance measures more suitable than others. The choice of a similarity measure is not trivial because it is necessary to find a relationship between the characteristics of the time series databases and the properties of the different distance measures. The common strategy is to experiment with a set of different distances and to choose one that is suitable one based on the obtained results. Due to the large number of available time series distance measures and the time complexity of many of them, this strategy is computationally very expensive and intractable in practice. The main objective is to train a multi-label classifier that, given a specific unlabeled time series database, will automatically select the most suitable distance measures from a set of candidates. The set of relevant features describes each time series database and which will support the choice of one distance measure over another.

4. Issues on Time Series Data Clustering Process

The time series data values are used to manage the transactional data with time and date information. The transactions are composed with binary, categorical, continuous and time information. The similarity measures are used to analyze the relationship between

the transactions. Normal similarity or distance measures are not suitable for the time series data values. The characteristics of time series data values are considered in the relationship analysis process. Shift, seasonal, trend, correlation, skewness and kurtosis features are extracted and analyzed for the time series data values. Different distance measures are employed to estimate the similarity values. Euclidean Distance (ED), Dynamic Time Warping (DTW), Edit Distance for Real Sequences (EDR), TQuest and Fourier Coefficient based Similarity Measure are adapted to estimate the relationship levels. The multi class classifiers are incorporated to select the suitable distance measure. Ensemble Classifier Chain (ECC) method and Random-k-labelsets classifier (RkL) methods are applied for the distance measure selection process. The clustering process is carried out with K-medoids clustering algorithm and agglomerative hierarchical algorithm. The following issues are identified from the current time series data clustering schemes.

- Temporal cost optimization is not supported
- Database uncertainty is not handled
- Limited scalability with high computational cost
- Data sampling and parameter reduction is not supported
- Distance measure and cluster centroid selection are not optimized

5. Optimal Distance Measures based Temporal Data Clustering

The temporal data clustering framework is constructed to partition the time series data values. The system supports partitional data clustering process on uncertain data values. Cluster centroid optimization scheme is integrated in the K-means clustering algorithm. The system is divided in to five major modules. They are Data preprocess, Similarity analysis, Distance Measure Selection Process, Centroid Optimized K-Means (COKM) Clustering Process and Distance and Centroid Optimized K-Means Clustering Process. The data preprocess module is designed to construct transaction matrix. Similarity analysis module is used to estimate the relationship between the transactions. The multi lebale classification operations are carried out under the distance measure selection process. The Euclidean distance is used in the Centroid Optimized K-Means (COKM) clustering scheme. The Distance and Centroid Optimized K-Means (DCOKM) clustering scheme is build with the Edit Distance for Real Sequences measure.

5.1. Data Preprocess

Sales transactions are maintained in data files. Uncertainty analysis is carried out to verify product count in each record. Transaction matrix group ups the bill and product details. Product code and product name are maintained in product list. The customer transactions form is used to fetch the customer data values from the database. The uncertainty analysis form is used show the uncertainty level for the transactions. The customer transactions are converted into transaction matrix. The product code and product name details are provided in product details form.

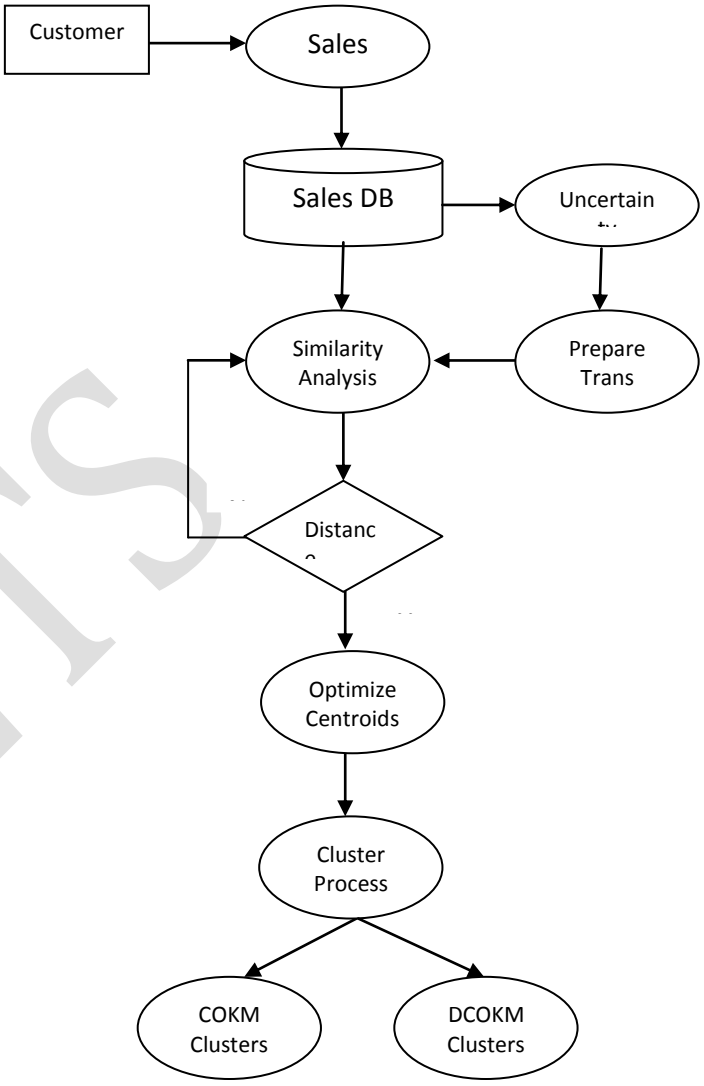

Figure No: 5.1. Temporal Data Clustering Framework

5.2. Similarity Analysis

Similarity measures are used to fetch the transaction relationships. Euclidean Distance (ED) and Edit Distance for Real Sequences (EDR) methods are used in the similarity analysis. The similarity analysis is used fetch the relationship between the transactions. Time series data values are compared in the similarity analysis process. Shift, seasonality, trend, correlation, skewness, kurtosis and auto correlation features are analyzed in the similarity estimation process.The similarity values are calculated and displayed in the same form.

5.3. Distance Measure Selection Process

The time series data clustering process uses different distance measures for transactional relationship analysis. The Euclidean Distance (ED), Dynamic Time Warping (DTW), Edit Distance for Real Sequences (EDR), TQuest andFourier Coefficient based Similarity Measure are adapted in the similarity analysis process. The multi class classifier is employed to select the distance measure. The Ensemble Classifier Chain (ECC) scheme and Random k-Labelest Classifier (RkL) scheme are used in the selection process. The Euclidean Distance (ED) and Edit Distance for Real Sequences (EDR) measures are listed in separate form. The distance measures are assigned for the clustering process.

5.4. Centroid Optimized K-Means (COKM) Clustering Process

Density and distribution analysis is performed on transaction matrix values. Transaction matrix is updated with distribution values. Centroid Optimized K-Means (COKM) clustering method is used for the clustering process. Centroid optimization scheme is used to improve the cluster accuracy levels. The density and distribution analysis is carried out to find out the transaction distribution levels. The Euclidean Distance (ED) measure is used in the clustering process. The clustering process is performed with cluster count collected from the user. The cluster results are produced in separate form. The Inter Cluster Distance (ICD) scheme is used to select optimal centroid values.

5.5. Distance and Centroid Optimized K-Means Clustering Process

Partitional clustering is performed using Distance and Centroid Optimized K-Means Clustering Process (DCOKM) schemes. The Edit Distance for Real Sequences (EDR) similarity is used with Distance and Centroid Optimized K-Means Clustering Process (DCOKM) schemes. Random Centroid selection scheme is replaced with optimal centroid selection model. Inter cluster distance analysis is applied in optimal centroid selection process. The clustering process is performed with cluster count collected from the user. The system uses the Inter Cluster Distance (ICD) based optimal centroid for clustering process. The clustering results are listed in separate form for the selected cluster name.

6. Conclusion and Future Enhancement

The time series data mining applications are build to analyze the transaction data with time information. The temporal data clustering framework is constructed to group the transactions with time series data values. The time series data values are compared using Euclidean Distance (ED), Dynamic Time Warping (DTW), Edit Distance for Real Sequences (EDR), TQuest and Fourier Coefficient based Similarity Measure. The distance measure selection process is carried out with the Classifier Chain algorithm (CC) and Random-k-labelsets classifier (RkL) The Euclidean Distance (ED) measure is used with Centroid Optimized K-Means (COKM) clustering scheme. The Edit Distance for Real sequences (EDR) distance measure is used in Distance and Centroid Optimized K-Means (DCOKM)cluster technique. The system is tested with customer sales transaction data values. The clustering framework achieves high cluster accuracy levels with minimum computation complexity levels. The system can be enhanced with the following features. The time series data comparison process is improved to handle time series data from different time zones. The distance measure selection process can be integrated to categorize the time series data values.

REFERENCES

[1] C. Aggarwal and K. Subbian, “Evolutionary network analysis: A survey,” ACM Comput. Surveys, vol. 47, no. 1, p. 10, 2014.

[2] J. Liu, L. Yuan, and J. Ye, “An efficient algorithm for a class of fused Lasso problems,” in Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2010, pp. 323–332.

[3] D. Mavroeidis and E. Marchiori, “Feature selection for k-means clustering stability: Theoretical analysis and an algorithm,” Data Min. Knowl. Discov., vol. 28, no. 4, pp. 918–960, 2014.

[4] F. Wang, H. Tong, and C.-Y. Lin, “Towards evolutionary nonnegative matrix factorization,” in Proc. 25th AAAI Conf. Artif. Intell., vol. 11, 2011, pp. 501–506.

[5] L. Wasserman, M. Azizyan, and A. Singh, “Feature selection for high-dimensional clustering,” arXiv preprint arXiv:1406.2240, 2014.

[6] M. Deodhar and J. Ghosh, “SCOAL: A framework for simultaneous co-clustering and learning from complex data,” ACM Trans. Knowl. Discov. Data, vol. 4, no. 3, pp. 11:1–11:31, 2010.

[7] K. S. Xu, M. Kliger, and A. O. Hero Iii, “Adaptive evolutionary clustering,” Data Min. Knowl. Discov., vol. 28, no. 2, pp. 304–336, 2014.

[8] C. Chen, J. Huang, L. He, and H. Li, “Preconditioning for accelerated iteratively reweighted least squares in structured sparsity reconstruction,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2013, pp. 2713–2720.

[9] H. H€ofling, “A path algorithm for the fused Lasso signal approximator,” J. Comput. Graph. Stat., vol. 19, no. 4, pp. 984–1006, 2010.

[10] J. Huang, S. Zhang, H. Li, and D. Metaxas, “Composite splitting algorithms for convex optimization,” Comput. Vis. Image Understanding, vol. 115, no. 12, pp. 1610–1622, 2011.

[11] M. Lee, H. Shen, J. Z. Huang, and J. S. Marron, “Biclustering via sparse singular value dec

[12] Rongjian Li, Wenlu Zhang, Yao Zhao, Zhenfeng Zhu and Shuiwang Ji, “Sparsity Learning Formulations for Mining Time-Varying Data”, IEEE Transactions on Knowledge and Data Engineering, Vol. 27, no. 5, May 2015.