International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

Similarity Measure Selection for Clustering Time Series Databases using Leading Activities

T. Karthikeyan
Assistant Professor, Department of CSE
Knowledge Institute of Technology
Salem, Tamilnadu, India

Dr. T. Sitamahalakshmi
Professor, Department of CSE
GITAM University
Visakhapatnam, Andhra Pradesh, India

***Abstract* - Data mining, the follow of examining massive pre-existing databases so as to get new data. The goal of data mining method is obtaining information from knowledge set and enhancing it to lucid structure for the further use. Data mining is considered to be the scrutinizing process of "knowledge discovery in databases" process, or KDD. A series of information points are indexed (tabulated or portrayed) in time order. Typically mentioning, a statistic takes an array of time at equally spaced points. So it's a sequence of discrete-time knowledge. Measuring apex of ocean currents, counting of sunspots and the daily end worth of the Dow-Jones Industrial Average are taken as the real time samples of statistic square. In this paper, the team tend to create a comparative study of all papers and to resolve the typical drawback of all the papers. The comparative study shows that it's terribly troublesome to investigate, study and to extract data from statistic knowledge sets. The main drawback is usage of different algorithm by every author. This paper concentrates on eliminating the drawbacks till the most accuracy is obtained. It's a laborious task to extract all data from the given set of the time series. This paper concentrates on the comparative study of different papers and proposes a typical solution for the drawbacks.**

Key Words: Data Mining, Time Series, Accuracy, Aggregation, Rating.

I. INTRODUCTION

The data mining is a method of interpreting data from disparate perceptions and consummating it into an aiding data that increases revenue, cut costs or in some cases both. Data processing software is an analytical tool for examining the data. It permits the user to explore information from various aspects and also allows reasoning or identifying. Technically, data mining is to find associations or similar patterns among diverse of fields in umpteen relative databases. The ultimate aim of the data mining is to garner evidence from an information set and transform it into a clear structure for superfluous use. In the raw analysis step, it comprises aspects of data management and information, pre-processing information, issues dealt with model, quality and reason, interest metrics, post-processing of discovered structures, visual image, and on-line change. Data processing contains five major elements: Extract, transform, and load dealing information on to the information warehouse system.

The main purpose of this is to store and manage the information in a third-dimensional information system. It also provides information access to business analysts and experts of data technology. It correlates the information by application software package. The information is exhibited in useful formats like graph or table. Classification and cluster are instances of machine learning in a supervised or unsupervised learning manner. The nuance between supervised and unsupervised learning is that the former distributes the instances to predefined categories whereas the later does not require it. In real valued statistical analysis, Continuous information, separate numeric information or separate symbolic information is applied. The space between two points of Euclidean space is known as the Euclidian metric the norm associated is named the Euclidian norm. The L2 norm or L2 distance is the generalized known term. The Euclidian distance between point's p and Q is that the length of the line phase connecting p and Q. To predict the label of unlabelled statistic a statistic classification is created to support the model. The statistic classification with R extracts and build features initially from statistic information and so applies prevailing sorting methods, like Support Vector Machine, k-NN, Neural Networks, regression and decision trees, to the basic feature set. The pragmatic observation of Statistic information has been done in the past few years and it depicts that dimensionality in statistic have a control on accuracy and performance. For instance, if the accuracy of rigid distance measures like Euclidian distance is directly proportional to the range of time series within the information. Besides, not all distance measures give sensible results only long time operation provides results. Consuming lot of time and economical multi-label classification are considered to be the major drawbacks as it cannot be used in large information set.

II. SURVEY WORK

A. Correlation Analysis Techniques

Data reconstruction or sophisticated method with irregular time samplings is provided by The Geo-scientific measurement. The linear interpolation techniques are compared with different approaches for analyzing the correlation functions, as Lomb- Scargle Fourier transformation and kernel based methods.

Low skewness and high skewness are the major components of the Root Mean Square Errors (RMSE). Lomb-Scargle technique applies the kernel methods for univariate result but the end result is the result is bivariate [7]. The kernel method uses the magnitude of the Strong bias in Autocorrelation function (ACF) and Cross Correlation function (CCF). In this paper, Cross Correlation functions used are similar to the common variability. The main advantage of Gaussian kernel is its reliability and affinity. The notable advantage of this is in large scale application to paleo data.

For regular and irregular sampling, The RMSE and bias schemes are employed using interpolation to synthetic records. The interpolation and Fast Fourier Transform (FFT) based routine increases RMSE and the absolute value of the bias four to seven times.

ACF estimation of sinusoidal signals is performed by the sinc -kernel. The linear interpolation possesses both pros and cons effect: ACF estimation withholds positive bias, and CCF estimation withholds negative bias.

In the cross correlation analysis of the paleo records, the interpolation and cross correlation consists of the kernel based lag zero cross correlation functions. A strongly negative bias is shown in the interpolation process with little persistence in the CCF estimation. The Gaussian kernel and the interpolation techniques in ACF estimation cause the problem in the analysis when sampling.

B. Decomposing Time Series Data

STL (Seasonal and Trend decomposition uses Loess) uses seasonal and remainder components for filtering and decomposing procedure of the time series.

The properties are its rapid computation, and long time. The features of STL are ranges specification of seasonal and smoothing, estimation of trends and non-distortion in behaviour data, ability to decompose the missing values and the period specification of multiple integers. The iterated weighed least squares is the method of STL which estimates the regression of the square fitting.

The time series of the STL, the linear operator is applied straight to the seasonal component.

C. Methods in Analysing Time-Series Data Mining

It is a hard task for the computers to extract all the possible knowledge from the given data which is possible by the human. Data mining is done for analysing various techniques or methods in time series [2]. Thus, it provides a good comprehension of the time series in data mining research field. Thus, various research papers have been discussed in this paper. More than the research, the algorithms proposed in these papers raises more question. Thus, the answer lies in the question.

D. Random k-Label Sets

The Random K-label sets (RAKEL) algorithm is an ensemble method used for multilabel classification. This ensemble method provides It is profound that this ensemble provides a better performance than the popular multi-label classification approach. Based on the space available, a new ensemble method for multi-label classification is proposed. Contrasting the popular Binary Relevance (BR) method and the LP classifier, the latter shows better performance than the previous one.

The inclusion of random nature of RAKEL affects the ensemble performance of the model[4]. In order to avoid this problem, coupling of RAKEL with an ensemble selection method will eventually lead to a better performance.

E. Classifier Chains for Multi- Label Classification

The binary relevance based methods are used in terms of scalability to large datasets. The chaining methods are used in terms of predictive performance and time complexity [3]. The empirical evaluation is broad range of multi-label datasets with a variety of evaluation metrics.

The classifier chains improve state-of-the-art methods, on large datasets. The advantage is the sophisticated current methods, in terms of time costs and the disadvantage is binary method in maintaining acceptable complexity.

F. Multi- Label Learning Algorithms

The emphasis on state-of-the-art multi –label learning algorithm is shown below. At First, the fundamentals on multi-label learning, they are formal definition and evaluation metrics. Second, eight representative multi-label learning algorithm and relevant analyses [8].Third, the many learning settings square measure on-line resources and open analysis issues on multi-label learning.

Paradigm systematization, learning algorithms and connected learning deals with progressive learning. Asymmetric label influences one label to the other in the inverse direction. Ensemble learning techniques ensures the ranking metrics in classification.

G. Distance Measures for Time Series Data

The experimental is re-implementing eight different time series representation and nine similarity measures and their variants, 38 time series data sets from a wide variety of application domains.

Our experiments were carried out on 38 diverse time series data sets which includes the tightness of low level bounding, pruning power and efficiency of different methods [6]. The accuracy level is measured using LCSS, EDR and EPR which are very close to DTW. The magnitude size 50 to 2,000 is the error rate that is reduced in DTW and Euclidean distance.

H. Ranking Methods

The paired and unpaired comparisons lead to variety of differences that may be positive or negative [5]. In the same way unpaired mortality and paired have the difference in height and self-fertilized plants of same pair. This paper depends on the knowledge of the tools and also particular domain.

III. PROBLEM IDENTIFICATION

Time series data clarification has been analysed in the past few years and the problem identifies in this area is that the dimensionality affects the accuracy of the time series. For example, Euclidean distances enhance the time series and the accuracy elevates. All the times series does not produce good results. The drawbacks are,

- Consumes more time
- Efficient multi-label classification is not discussed
- Not applicable for large data set

IV. PROPOSED SYSTEM

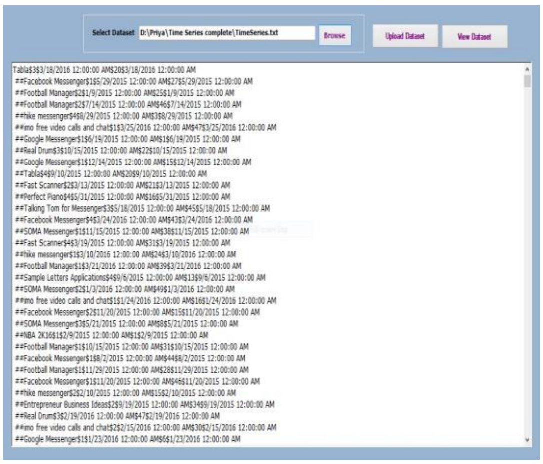
A set of relevant features are proposed to describe each database this will not only support the time series database but also the characteristics in predicting the information from the suitable distance. Considering distance measures, parameter settings and clustering algorithms clustering of database occurs. For training the proposed multi-label classifier two algorithms have been chosen. But any algorithm can be used to work with frame work in a direct manner. Binary classification is the major problem decomposes the multi-label problem into an ordered chain. We have used synthetic data sets. It uses random classifier.

The list of modules proposed,

- Data Processing
- Time Series Database Characterization
- Multi-label Classifier
- Evidence aggregation
- Classification-app recommendation

A. Data Processing

The dataset hold enormous amount of data. The data can be structured or unstructured. If the data is unstructured each phase is analysed with the parameters used in the transaction. In Thus, the unstructured dataset is converted into structure dataset. Data processing can be done in manual, automatic and electronic processing. Manual processing is of antique style which involves the process of garnering the information periodically with analysis and presentation, like accounting and census exercises. Electronic processing is the widespread, trendy technique of grouping, manipulating, analysing and presenting information and data. Electronic processing is additionally referred to as processed processing.



Select Dataset: D:\Priya\Time Series complete\TimeSeries.txt

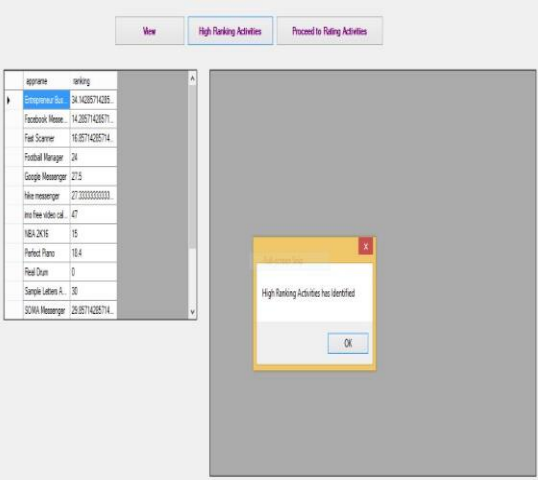
```
Tabla$3$3/18/2016 12:00:00 AM$20$3/18/2016 12:00:00 AM
 ##Facebook Messenger$1$5/29/2015 12:00:00 AM$27$5/29/2015 12:00:00 AM
 ##Football Manager$2$1/9/2015 12:00:00 AM$25$1/9/2015 12:00:00 AM
 ##Football Manager$2$7/14/2015 12:00:00 AM$46$7/14/2015 12:00:00 AM
 ##hike messenger$4$8/29/2015 12:00:00 AM$3$8/29/2015 12:00:00 AM
 ##imo free video calls and chat$1$3/25/2016 12:00:00 AM$47$3/25/2016 12:00:00 AM
 ##Google Messenger$1$6/19/2015 12:00:00 AM$1$6/19/2015 12:00:00 AM
 ##Real Drum$3$10/15/2015 12:00:00 AM$22$10/15/2015 12:00:00 AM
 ##Google Messenger$1$12/14/2015 12:00:00 AM$15$12/14/2015 12:00:00 AM
 ##Tabla$4$9/10/2015 12:00:00 AM$20$9/10/2015 12:00:00 AM
 ##Fast Scanner$2$3/13/2015 12:00:00 AM$21$3/13/2015 12:00:00 AM
 ##Perfect Piano$4$5/31/2015 12:00:00 AM$16$5/31/2015 12:00:00 AM
 ##Talking Tom for Messenger$3$5/18/2015 12:00:00 AM$45$5/18/2015 12:00:00 AM
 ##Facebook Messenger$4$3/24/2016 12:00:00 AM$43$3/24/2016 12:00:00 AM
 ##SOMA Messenger$1$11/15/2015 12:00:00 AM$38$11/15/2015 12:00:00 AM
 ##Fast Scanner$4$3/19/2015 12:00:00 AM$31$3/19/2015 12:00:00 AM
 ##hike messenger$1$3/10/2016 12:00:00 AM$24$3/10/2016 12:00:00 AM
 ##Football Manager$1$3/21/2016 12:00:00 AM$39$3/21/2016 12:00:00 AM
 ##Sample Letters Applications$4$9/6/2015 12:00:00 AM$13$9/6/2015 12:00:00 AM
 ##SOMA Messenger$2$1/3/2016 12:00:00 AM$49$1/3/2016 12:00:00 AM
 ##imo free video calls and chat$1$1/24/2016 12:00:00 AM$16$1/24/2016 12:00:00 AM
 ##Facebook Messenger$2$11/20/2015 12:00:00 AM$15$11/20/2015 12:00:00 AM
 ##SOMA Messenger$3$5/21/2015 12:00:00 AM$8$5/21/2015 12:00:00 AM
 ##NBA 2K16$1$2/9/2015 12:00:00 AM$1$2/9/2015 12:00:00 AM
 ##Football Manager$1$10/15/2015 12:00:00 AM$31$10/15/2015 12:00:00 AM
 ##Facebook Messenger$1$8/2/2015 12:00:00 AM$44$8/2/2015 12:00:00 AM
 ##Football Manager$1$11/29/2015 12:00:00 AM$28$11/29/2015 12:00:00 AM
 ##Facebook Messenger$1$11/20/2015 12:00:00 AM$46$11/20/2015 12:00:00 AM
 ##hike messenger$2$2/10/2015 12:00:00 AM$15$2/10/2015 12:00:00 AM
 ##Entrepreneur Business Ideas$2$9/19/2015 12:00:00 AM$34$9/19/2015 12:00:00 AM
 ##Real Drum$3$2/19/2016 12:00:00 AM$47$2/19/2016 12:00:00 AM
 ##imo free video calls and chat$2$2/15/2016 12:00:00 AM$30$2/15/2016 12:00:00 AM
 ##Google Messenger$1$1/23/2016 12:00:00 AM$6$1/23/2016 12:00:00 AM
```

Fig.1.Time Series Data set Value

B. Time Series Database Characterization

To predict information of a distance a set of characteristics of time series is used. We create synthetic dataset which is based on ranking, rating, review based evidences which are all in time series. There are two main steps in characterization.

- ➢ Discovering the records of apps
- ➢ Constructing leading sessions by merging with events.



appname	ranking
Entrepreneur Bus...	34.14285714285...
Facebook Messe...	14.2857142857...
Fast Scanner	16.85714285714...
Football Manager	24
Google Messenger	27.5
hike messenger	27.3333333333...
imo free video cal...	47
NBA 2K16	15
Perfect Piano	18.4
Real Drum	0
Sample Letters A...	30
SOMA Messenger	29.85714285714...

High Ranking Activities has Identified

Fig. 2.High Ranking Activities for App

C. Multilabel Classifier

In this multi label classifier, the characteristics used automatically choose the suitable distance from the evidence based on the time.

- ➢ Ranking Based evidences
- ➢ Rating Based Evidences
- ➢ Review Based Evidences

In machine learning, multi-label classification possesses a predominant problem where the variants assigned must have a multiple label assigned for every variant. Multiclass and multi-label should not be confused in classification. It normally has two drawbacks. Basically, multi-label learning is has a drawback of recognizing the binary vectors y and map inputs x in the classification. There are 2 ways in multi-label classification problem,

- Problem transformation ways
- Formula adaptation ways

Problem demodulation ways converts multi-label into a herd of binary classification by single class classifiers. Formula adaptation ways adapt to the algorithm to directly perform multi-label classification.

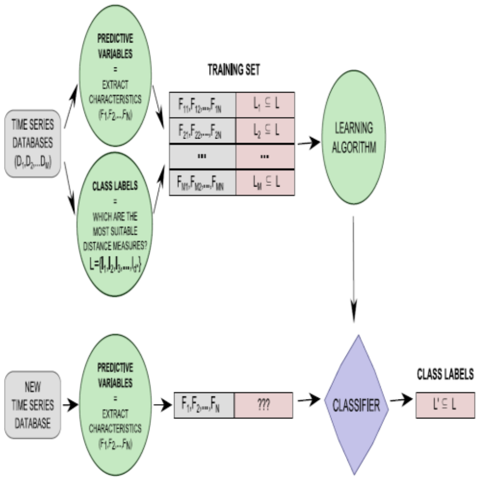

TIME SERIES DATABASES (D_1,D_2...D_M)
PREDICTIVE VARIABLES = EXTRACT CHARACTERISTICS (F1,F2...FN)
CLASS LABELS = WHICH ARE THE MOST SUITABLE DISTANCE MEASURES? L={l_1,l_2,l_3,...,$l_{|L|}$}
TRAINING SET
F_{11},F_{12},...,F_{1N}
$L_1 \subseteq L$
F_{21},F_{22},...,F_{2N}
$L_2 \subseteq L$
...
...
F_{M1},F_{M2},...,F_{MN}
$L_M \subseteq L$
LEARNING ALGORITHM
NEW TIME SERIES DATABASE
PREDICTIVE VARIABLES = EXTRACT CHARACTERISTICS (F1,F2...FN)
F_1,F_2,...,F_N
???
CLASSIFIER
CLASS LABELS
$L' \subseteq L$

Fig. 3.Multi label classifier

D. Evidence aggregation

Ranking the fraud detection is the tedious challenge after extracting the fraud evidences. Many models such as permutation based models, score based models and Dempster-Shafer rules are proposed to segregate and rank them in order. However, these ranking exposes the fake ranking for all users worldwide. So this is not a proper method for ranking the fake apps. Methods based on supervised learning techniques are found to be a hard thing to exploit. Instead, an unsupervised approach has been prepared to sort out the similarity. The combined evidences provides the best and the fraudulent application details.

E. Classification-App Recommendation

It is very helpful to the mobile user to choose best apps and to avoid fraud apps before downloading. This module is used to classify the fraud app among the various app by the time series dataset characterization. It compares the evidence aggregated result with the leading session better apps.

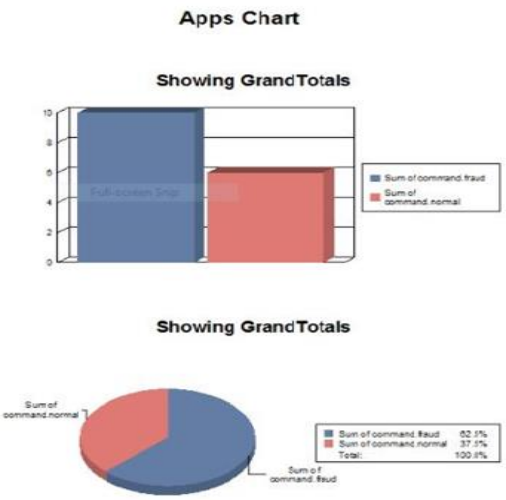


Fig.4. Leading Activities for App Chart

V. CONCLUSION

For clustering of time series database, automatic measure selection of multi-label classifier has been proposed. The database receives a set of characteristics as input and the output will be in the suitable distance measure. This method is enhances the performance level which simplifies the distance measure using time series clustering task. An evaluation method is proposed to evaluate the clustering task. No method is proposed before as stated in this paper.

VI. FUTURE ENHANCEMENT

In future the research can include new distance measures in the proposed framework. Additionally an extensive research can be made by adding new features that would describe the time series which is not included in this paper. To improvise the research, the features stated in this paper can be taken into account.

Next proposal that can be considered on hand is optimization of the temporal costs in association with the characteristic calculation. Inclusive features that are introduced in this study are the shift which is quite expensive to calculate which could cause inconvenience particularly in large databases. In the same way, decreasing the parameters associated to the characteristics could improve the applicability of the proposal.

REFERENCES

[1] Cleveland, R. B, Cleveland, W .S, McRae, J .E and Terpenning. I. (1990) “STL: A Seasonal-trend decomposition procedure
based\on loess,” J. Official Statist., vol. 6, no. 1, pp. 3–73.

[2] Esling, P. and Agon, C.(2012) “Time-series data mining,” ACM Comput. Surveys, vol. 45, no. 1, pp. 1–34.

[3] Holmes, G. and Frank, E. (2011)“Classifier chains for multi-label classification,” Mach. Learning, vol. 85, pp. 333–359.

[4] Tsoumakas, G. and Vlahavas, I. (2007)“Random k-Labelsets: An ensemble method for multilabel classification,” in Proc. 18th Eur. Conf. Mach. Learning, pp. 406–417.

[5] F. Wilcoxon, “Individual comparisons by ranking methods,” Bio-metrics, vol. 1, no. 6, pp. 80–83, 1945.

[6] M. Hubert and E. Vandervieren, “An adjusted boxplot for skewed distributions,” Comput. Statist. Data Anal., vol. 52, no. 12.

[7] M. M. Breunig, H.-p. Kriegel, and R. T. Ng, “LOF: Identifying density-based local outliers,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2000, pp. 93–104.

[8] M.-L. Zhang and Z.-H. Zhou, “A review on multi-label learning algorithms,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[9] L.Chen, M. T. Ozsu, and V. Oria, “Robust and fast similarity search for moving object trajectories,” in Proc. ACM SIGMOD Int.Conf. Manage. Data, 2005, pp. 491–502.

[10] J.Aßfalg, H.-p. Kriegel, P. Kroger, P. Kunath, A. Pryakhin, and M.Renz, “Similarity search on time series based on threshold queries,” in Proc. 10th Int. Conf. Adv. Database Technol., 2006, pp. 276–294