ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.7

Similarity Measure Selection for Clustering Stock Market Time Series Databases

Dr.S.Radhimeenakshi[1], K.Latha[2]
Associate Professor[1], Research Scholar[2]
Department of Computer Science
Tiruppur Kumaran College for Women, Tirupur, Tamilnadu, India

Abstract:
In the past few years, clustering has become a popular task associated with time series. The choice of a suitable distance measure is crucial to the clustering method and, specified the vast number of distance measures for time series available in the literature and their diverse characteristics, this selection is not uncomplicated With the objective of simplifying this task, we can use a multi-label classification framework that provides the means to automatically select the most suitable distance measures for a time series database. This thinker is based on a novel collection of characteristics that describe the main features of the time series databases and provide the predictive information necessary to discriminate between a set of distance measures. This Context provides a survey of some of the approaches for time series database clustering.

Keywords: Time series, distance measures, clustering, multi-label, database clustering, framework.

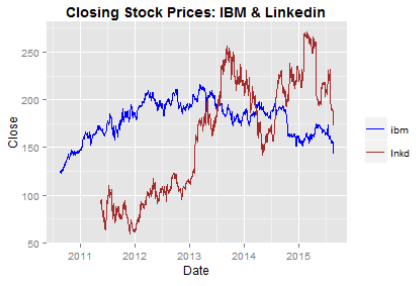
1. INTRODUCTION

1.1. TIME SERIES DATA MINING:

A Time series is a set of explanation where each one being recorded at a specific time. It is of two types, a discrete-time series is the first one in which the set of times at which observations are made as a discrete set, for example the observations are made at fixed time intervals. Second continuous time-series are observed where observations are recorded continuously over some time interval. Major time-series related tasks include query by content, anomaly detection, motif discovery, prediction, clustering, classification and segmentation. Time-series data mining unveils several facets of complexity. The most important problems are similarity measures, data representations and indexing methods. This work reviewed some of the time-series data mining tasks.

2. TIME SERIES DATA ANALYSIS APPLICATION:

A time series is a series of data point indexed (or listed or graphed) in time arrange. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Cases of time arrangement are statures of sea tides, number of sunspots, and the day by day finishing up estimation of the Dow Jones Industrial Average. Time arrangement is habitually plotted by means of line graphs. Time arrangement are second-deliver climate estimating, flag preparing, econometrics, design acknowledgment, scientific fund, insights, smart transport and direction guaging, electroencephalography, control building, cosmology, seismic tremor expectation, correspondences building, and fundamentally in any space of down to earth science and building which includes fleeting estimations. Time arrangement investigation involves techniques for examining time arrangement information keeping in mind the end goal to remove significant measurements and different qualities of the information. Time arrangement guaging is the utilization of a model to foresee future esteems in view of already watched esteems. While relapse investigation is regularly utilized so as to test hypotheses that the present estimations of at least one free time arrangement influence the present estimation of some other time arrangement, this kind of examination is called "time arrangement investigation", which concentrates on looking at estimations of a solitary time arrangement or various ward time arrangement at various focuses in time. Time series data have a natural temporal ordering. This makes time series analysis different from cross-sectional studies, in which there is no natural ordering of the observations (e.g. explaining people's wages by reference to their respective education levels, where the individuals' data could be entered in any order). Time series analysis is also separate from spatial data analysis where the observations typically relate to geographical locations (e.g. accounting for house prices by the location as well as the intrinsic characteristics of the houses). A stochastic model for a time series will normally reflect the fact that observations close together in time will be more closely related than observations further apart. In addition, time series models will often make use of the natural one-way ordering of time so that values for a given period will be expressed as deriving in some way from precedent values, rather than from future values.



3. TIME SERIES BASED STOCK MARKET ANALYSIS:

The stock market can be viewed as a exacting data mining and artificial intelligence problem. The movement in the stock exchange depends on capital gains and losses and most people believe the stock market unpredictable and irregular. on the other hand, patterns that agree to the prediction of some movements can be construct. Stock market analysis deals with the revise of these patterns. It uses distinctive techniques and strategies, mostly automatic that trigger buying and selling orders depending on different decision making algorithms. It can be measured as a bright treatment of past and present economic data in order to predict the stock market future performance. Therefore it can be viewed as an artificial intelligence problem in the data mining field. It is to construct and evaluate these investment strategies in order to predict future stock exchanges. Firstly, data mining approaches are used to evaluate past stock prices and acquire useful knowledge through the calculation of financial indicators. Transformed data are then confidential using decision trees obtained through the application of Artificial Intelligence strategies. Finally, the different decision trees are analyzed and evaluated, showing accuracy rates and emphasizing total profit related to capital gains.

3.1. STOCK MARKET PREDICTION

Stock market investment is an area that has been gaining long-term attention by financial institutions, individuals and multiple research communities. Judging a favorable investment decision on the plethora of available stocks in the market is a tiresome and challenging task. There is considerable amount of uncertainty about the nature of returns and hence poses difficulty in the decision-making process associated with selection of securities. There is a need to strike the right balance between expected return (maximize) and associated risk (minimize) There are multiple analytical methodologies employed for decision making in stock exchanges, which could be broadly categorized into two groups viz. Technical analysis and Fundamental analysis. Fuzzy Expert Systems as well as Artificial Neural Networks were employed to analyse the stock market and measure the attractiveness of the participating companies. it have designed a decision support system for projecting buying or selling decisions utilizing principles of fundamental analysis together with considering technical indicators. Primarily these systems were developed for novice investors to aid in making subjective judgments regarding stock selection as per their individual .The opinion of an expert is derived out of his experience gained by analyzing stock features over a period of time. The two key components of credibility of a recommendation identified the majority of researchers are trustworthiness and expertise in order to counter the uncertainty a Fuzzy Expert System will be presented. Another way of getting insight into what investors and traders thinking about a particular stock is carry out a sentiment analysis of Twitter data (tweets) of investors. Investors also connect with one another to discuss trade, invest, learn and share knowledge across the network. The analysis of the investor's network thus formed could provide insight into the wisdom of the crowds to help one make smarter investment decisions. However in such networks the trust between individuals cannot be fully depended and the relationships could be falsely built. If one knows that his tweets are followed then he can put fake tweets and false information to influence our investment decision, which is considered as the main problem of social network analysis. Discovering trust in relationships among entities of a social network which leads to a trust-based social network is a promising solution to this problem. This can be further improved by incorporating an expert opinion as the trusted expert advice will lead to better results. Attempts have been made to develop social network of financial experts based on their publicly listed portfolios for further analysis to recommend an appropriate portfolio to a novice investor. However classifying the type of knowledge that different experts have is a challenging problem. Apart from this the type of knowledge, particularly tacit knowledge gained through experience and learning over time is hard to be coded and also people are not often aware of the knowledge they possess or its value for others. All these make the task of finding trustworthy experts for an investment decision more complicated and challenging. With the objective of developing a trust based investment relationship, we propose a social network approach making use of Mutual Fund Investment Portfolio. It is generally believed that pursuing an investment decision of a trustworthy mutual fund is less risky than seeking advice from an individual expert. We can also examine the credibility of investment behavior and stock holding patterns of the mutual fund in real time and a stock recommendation system, namely, trust based stock recommendation system, showing the leading stocks appropriate for investment can be designed. Such system can also show the investment price range by analysing the mutual fund transaction in the stock market reducing the overall risk and increasing the profitability . In order to demonstrate the acceptability and reliability of the proposed methodology, we have carried out a detailed analysis of this model on CRISIL-1 rated Indian Mutual Funds. The results of this analysis strongly support the validity of the proposed portfolio recommendation model.

4. LITERATURE REVIEW

CLUSTERING OF TIME SERIES DATA:

T.Warren Liao [1] has proposed summarizes previous works that investigated the clustering of time series data in various application domains. The basics of time series clustering are presented, including general-purpose clustering algorithms commonly used in time series clustering studies, the criteria for evaluating the performance of the clustering results, and the events to establish the similarity/dissimilarity between two time series being compared, either in the forms of raw data, extracted features, or some model parameters. The past researches are organized into three groups depending upon whether they work directly with the raw data either in the time or frequency domain, indirectly with features extracted from the raw data, or indirectly with models built from the raw data. The uniqueness and limitation of previous research are discussed and several possible topics for future research are identified. Moreover, the areas that time series clustering have been applied to are also summarized, including the sources of data used. It is hoped that this review will serve as the steppingstone for those interested in advancing this area of research.

4.1. EXPERIMENTAL COMPARISON OF REPRESENTATION METHODS AND DISTANCE MEASURES FOR TIME SERIES DATA: Xiaoyue Wang, Abdullah Mueen..en [2] has proposed the research efforts in this context have focused on introducing new representation methods

for dimensionality reduction or novel similarity events for the fundamental data. In the immeasurable best part of cases, each person work introducing an exacting method has made specific claims and aside from the occasional theoretical justifications, provided quantitative experimental observations. However, for the most part, the comparative aspects of these experiments were too narrowly focused on demonstrating the benefits of the proposed methods over some of the previously introduced ones. In categorize to give a complete validation, we conducted a general experimental study re-implementing eight different time series representations and nine similarity measures and their variants, and trying their efficiency on 38 moments in time series data sets from a wide range of application domains. In this editorial, we give a summary of these different techniques and present our relative experimental findings regarding their effectiveness.

4.2. COMPARISON OF CORRELATION ANALYSIS TECHNIQUES FOR IRREGULAR SAMPLED TIME SERIES

K. Rehfeld, N. Marwan..en [3] has proposed the linear interpolation technique and different approaches for analyzing the correlation functions and persistence of irregularly sampled time series, as Lomb-Scargle Fourier transformation and kernel based methods. In a systematic standard test we consider the presentation of these techniques. Every part of methods has equivalent root mean square errors (RMSEs) for low skewness of the inter-observation time circulation. For high skewness, very irregular data, interruption bias and RMSE increase strongly. We find a 40% lower RMSE for the lag-1 Auto Correlation Function (ACF) for the Gaussian kernel method vs The linear interruption scheme, in the study of extremely asymmetrical time series. For the Cross Correlation Function (CCF) the RMSE is then lower by 60 %. The purpose of the Lomb-Scargle technique give results similar to the kernel methods for the univariate, but not as good as results in the bivariate case. Mainly the high regularity components of the signal, where classical methods show a strong bias in ACF and CCF magnitude, are preserved when using the kernel methods.

4.3. AN ADJUSTED BOXPLOT FOR SKEWED DISTRIBUTIONS

M. Hubert, E. Vandervieren [4] has proposed the boxplot is a very popular graphical tool to visualize the distribution of continuous uni-modal data. It shows information about the position, spread, skewness as well as the tails of the data. Though, when the data are skewed, frequently many points exceed the whiskers and are often incorrectly declared as outliers. An adjustment of the box plot is presented that includes a robust measure of skewness in the determination of the whiskers. This results in a more accurate representation of the data and of possible outliers. Consequently, this is an adjusted boxplot can also be used as a fast and automatic outlier detection tool without making any parametric assumption about the distribution of the bulk of the data. Several examples and simulation results show the advantages of this new procedure.

4.4. LOF: IDENTIFYING DENSITY-BASED LOCAL OUTLIERS Markus M. Breunig, Hans-Peter Kriegel..en[5] has proposed many KDD applications, such as detecting criminal activities in E-commerce, finding the rare instances or the outliers, can be more interesting than finding the common patterns. Presented work in outlier recognition regards being an outlier as a binary property. In this work, we contend that for many scenarios, it is more meaningful to assign an each object a *degree* of being an outlier. This level is called the *Local Outlier Factor* (LOF) of an object. It is *local* in that the level depends on how inaccessible the object is with respect to the surrounding neighborhood. We give a full formal analysis presentation that LOF enjoys many popular properties. Using real world datasets, we demonstrate that LOF can be used to find outliers which appear to be meaningful, but can otherwise not be identified with existing approaches as a final point, a attentive performance evaluation of our algorithm confirms we show that our approach of finding local outliers can be useful.

5. CLASSIFIER CHAINS FOR MULTI-LABEL CLASSIFICATION

Jesse Read, Bernhard Pfahringer..en [6] has planned the widely known binary significance method for multi-label classification, which considers each label as an autonomous binary problem, has often been over- looked in the writing due to the supposed inadequacy of not directly modeling label correlations. Most current methods invest considerable complexity to model interdependencies between labels. This employment shows that binary relevance-based methods have much to offer, and that high analytical presentation can be obtained without impeding scalability to large datasets. We exemplify this with a novel classifier chains method that can model label correlations while maintaining acceptable computational complexity. We extend this approach further in an ensemble framework. A wide-ranging experimental evaluation covers a broad range of multi-label datasets with a variety of evaluation metrics. The results illustrate the competitiveness of the chaining method against related and state-of-the-art methods, both in terms of predictive performance and time complexity.

5.1. A REVIEW ON MULTI-LABEL LEARNING ALGORITHMS

Min-Ling Zhang, Zhi-Hua Zhou[7] has proposed the past decade, significant amount of progresses have been made towards this rising machine learning paradigm. This work aims to provide a timely review on this area with importance on state-of-the-art multi-label learning algorithms. Firstly, basics on multi-label knowledge including formal description and evaluation metrics are given. Secondly and primarily, eight representative multi-label learning algorithms are scrutinized under common notations with relevant analyses and discussions. Thirdly, several related learning settings are briefly summarized. As an execution, online possessions and open research problems on multi-label learning are outlined for position purposes.

5.2. RANDOM *K*-LABELSETS: AN ENSEMBLE METHOD FOR MULTILABEL CLASSIFICATION

Grigorios Tsoumakas, Ioannis Vlahavas[8] has proposed in this effort proposes an company technique for multilabel classification. The Random *k*-label sets (RAKEL) algorithm constructs each member of the ensemble by considering a small random subset of labels and learning a single-label classifier for the prediction of each element in the power set of this subset. In this way, the proposed algorithm aims to take into account label

correlations using single-label classifiers that are applied on subtasks with manageable number of labels and adequate number of examples per label. Investigational results on common multi-label domains connecting protein, document and scene classification show that better presentation can be achieved compared to popular multilabel organization approaches.

5.3. A COMPLEXITY-INVARIANT DISTANCE MEASURE FOR TIME SERIES

Gustavo E.A.P.A. Batista**,** Xiaoyue Wang..en[9] has proposed The ubiquity of time series data across almost all human endeavors has produced a great interest in time series data mining in the last decade. While there is a plethora of classification algorithms that can be applied to time series, all of the current empirical evidence suggests that simple nearest neighbor classification is exceptionally difficult to beat. The choice of distance measure used by the nearest neighbor algorithm depends on the invariance's required by the domain. In this work we make a surprising claim. There is an invariance that the community has missed, *complexity invariance*. Intuitively, the problem is that in many domains the different classes may have different complexities, and pairs of complex objects, even those which subjectively may seem very similar to the human eye, tend to be additional apart under present distance measures than pairs of simple objects. This fact introduces errors in nearest neighbor classification, where complex objects are incorrectly assigned to a simpler class.

5.4. SIMILARITY SEARCH ON TIME SERIES BASED ON THRESHOLD QUERIES

Johannes Aßfalg**,** Hans-Peter Kriegel**..**en [10] has proposed the most prominent work has focused on similarity search considering either complete time series or similarity according to subsequences of time series. For many domains like financial analysis, medicine, environmental meteorology, or environmental observation, the detection of temporal dependencies between different time series is very important. In contrast to traditional approaches which consider the course of the time series for the purpose of matching, coarse trend information about the time series could be sufficient to solve the above mentioned problem. In particular, temporal dependencies in time series can be detected by determining the points of time at which the time series exceeds a specific threshold. In this job, we begin the book thought of threshold queries in time series databases which statement those time series greater than a user-defined query threshold at similar time frames compared to the query time series. We present a new resourceful access method which uses the fact that only partial information of the time series is necessary at query time. The performance of our solution is demonstrated by an extensive investigational evaluation on real world and an artificial time series data.

5.5. MEKA: A MULTI-LABEL/MULTI-TARGET EXTENSION TO WEKA Jesse Read [11] has considered the Multi-label classification has quickly involved interest in the machine learning writing, and there are now a large number and important variety of methods for this type of learning. We present Meka: an open-source Java structure based on the well-known Weka library. Meka provides interfaces to make possible practical application, and a wealth of multi-label classifiers, assessment metrics, and tools for multi-label experiments and expansion. It supports multi-label and multi-target data, as well as in incremental and semi-supervised contexts.

5.6. USING DYNAMIC TIME WARPING TO FIND ATTERNS IN TIME SERIES

Donald J. Bemdt, James Clifford [12] has proposed the Knowledge discovery in databases presents many interesting challenges within the content of providing computer tools for exploring large data archives. Electronic data repositories are growing quickly and contain data from commercial, scientific, and other domains. Much of this data is inherently temporal, such as stock prices or NASA telemetry data. Detect £bug patterns in such data streams or time series is an important knowledge discovery task. This work describes some primary experiments with a dynamic prograrning approach to the problem. The pattern detection algorithm is based on the dynamic time warping technique used in the speech recognition field. Keywords: dynamic programming, dynamic time warping, knowledge discovery, pattern analysis, time series.

6. CONCLUSIONS AND FUTUREWORK

In this work, a multi-label classifier for the automatic similarity measure selection has been proposed for the task of clustering time series databases. The classifier receives a set of characteristics that describe the database as input and returns the set of most suitable distance measures from a set of candidates. The positive results obtained in the experimentation for various multi-label classifications. This tool is useful to simplify the distance measure selection process, crucial to the time series database clustering task. The first obvious future research direction is to include new distance measures in the proposed framework. In this line, a more extensive study could be performed introducing new features that would describe other aspects of the time series databases that have not been considered in this work. For this purpose, some of the features presented in could be considered. Another proposal for future work includes an optimization of the temporal costs associated with the calculation of the characteristics. Some of the features introduced in this study, such as the shift, are computationally quite expensive to calculate, which could be an inconvenience when working with particularly large databases. Since only means, medians, standard deviations and other general statistics are calculated; strategies such as sampling the time series database could be applied to reduce this computational cost. In the same line, reducing the number of parameters associated to the characteristics could also improve the applicability of the proposal. Finally, some insights into the definition of the parameters of the distance measures have been included throughout the work, but no extended experimentation has been carried out on this topic. Studying the relationship between the characteristics of the databases and the parameters that define each distance could be useful to simplify the selection of a distance measure even more.

7. REFERENCES

[1]. T. W. Liao, "Clustering of time series data: A survey," Pattern Recog., vol. 38, no. 11, pp. 1857–1874, Nov. 2005.

[2]. X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh, "Experimental comparison of representation methods and distance measures for time series data," Data Mining Knowl. Discovery, vol. 26, no. 2, pp. 275–309, Feb. 2012.

[3]. K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths, "Comparison of correlation analysis techniques for irregularly sampled time series," Nonlinear Processes Geophysics, vol. 18, no. 3, pp. 389–404, Jun. 2011.

[4]. M. Hubert and E. Vandervieren, "An adjusted boxplot for skewed distributions," Comput. Statist. Data Anal., vol. 52, no. 12, pp. 5186–5201, Aug. 2008.

[5]. 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.

[6]. J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chainsfor multi-label classification," Mach. Learning, vol. 85, pp. 333–359,2011.

[7]. 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

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

[9]. G. E. Batista, X. Wang, and E. J. Keogh, "A complexity-invariant distance measure for time series," in Proc. SDM, 2011, pp. 699–710.

[10]. J. Aßfalg, H.-p. Kriegel, P. Kr€oger, 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.

[11]. J. Read. (2012). MEKA: A multi-label extension to WEKA [Online].Available: http://meka.sourceforge.net/

[12]. D. J. Berndt and J. Clifford, "Using dynamic time warping to findpatterns in time series," in Proc. KDD Workshop, 1994, pp. 359–370.