RESEARCH ARTICLE OPEN ACCESS

Interpolation of Stock market Data with Fuzzy Conception Using Weka Tool

Priti Choudhary , Vinod Rampure
[1]*M.Tech Scholar I.T. MITRC Alwar ,India*
[2]*Assistant professor CSE Dept. MITRCGECA Alwar, India*

ABSTRACT

Progressing growth of IT has brought rapid technological advancement. Technologies are getting advance at an exponential rate and hence massive amount of data is emerging at very enormous rate in different sector. So there are lots of baselines for researcher to roadmap their strategy for technological improvement. Huge amount of data i.e. terabytes of data are carried over computer networks to and from organization working in the field of business, engineering and science. Many approaches based on mathematical model were suggested for dredging association rule but they were complex for users. Our work contemplated an algorithm for interpolating Stock Market data using fuzzy data dredging through which fuzzy association rule can be induced for Stock series. Our work proposes the algorithm in which each fuzzy item has its own predefined minimum support count. Time series data can be any sequence data which has some trend or pattern in it. It may be either stock market data, climatic observed data, data observed from medical equipments. Our work also measures the data dispersion in time series data i.e. stock market data used here. It shows the deviation of the stock prices from the mean of stock price data points taken over a period of time which help the investors to decide whether to buy or sell their shares or products. Risk associated with particular share can also be predicted by understanding the obtained curve in the experiment. We have implemented the contemplated work in WEKA tool to get more accurate and efficient result along with visualization. Basically we are predicting how data are interpreted and predicted with accuracy in stock market using this effective tool.

Index Terms: WEKA, fuzzy association rule

I. INTRODUCTION

An enormous amount of data is emerging in the field of science, medical and many other areas due to the rapid development in computerization and digitalization techniques. These data may provide a great resource for knowledge extraction and decision support. In order to realize, analyze, and eventually make efficient use of the huge amount of data, a multidisciplinary approach is needed to meet the challenge. Traditionally we were forceful to depend on file management System and manual work. But now we are moving towards a new age called ―data age‖ e.g.-online shopping, railway ticket booking etc. are getting dependent on computer which contain a large amount of data. Several databases and data warehouses are built to capture the data. So a versatile and powerful tool is required to transform data into the valuable information. This led to the birth of data mining or data dredging. Many applications are based on data dredging e.g. – business intelligence, search engine etc. Business intelligence needs the past details and predict for future on the basis of the calculation. As gold mining is the exploration for chunks of gold, so data mining also known as data dredging is the exploration for chunks of information. In time series data dredging, these chunks are called as events. As gold is buried in the ground, chunks of information are masked in data. Energetic efforts have been done in designing efficient mechanisms for extracting information and rules from large databases. Data mining means the application of vital procedures for recognizing effective, rational, potentially useful, and previously unknown trends in large databases. Depending on the types of databases processed, knowledge discovery from database approaches may be divided as transactional databases, temporal databases, and relational databases. Secondly, depending on the classes of information derived, knowledge discovery from database approaches may be divided as inducing association rules, sequential patterns, clustering rules and classification rules. Inducing association rules in transaction databases is most common application in data mining. An association rule can be symbolized in the form P → Q, where P and are item sets, in such a way that the presence of P in a transaction will signify on the presence of Q. Two measures, support and confidence, are evaluated to determine whether a rule should be kept. Support and confidence, are evaluated to determine whether a rule should be kept. The support of a rule is the percentage of the transactions which include all the items in P and Q.

1.1.1 Association Rule

For a given transaction database T, an association rule is an expression of the for X → Y, Where X and Y are subsets of A and X → Y, holds with confidence c, if c % of transactions in T that support X also support Y, The rule X →Y has support s, if s% of transaction in T that support X ∪ Y. Support and confidence are two basic measures for association rule. Since database is large and users are concerned only about frequently appearing items hence in order to identify the set of association rule having reasonable support and confidence there are two user specified thresholds i.e. minimal support S and minimal confidence C .Thus X → Y is an association rule with confidence c if c ≥ C and support s if s ≥ S. 1.2 Data Mining or Data Dredging Tasks – It is divided into four categories. They are discussed as follows:-

- **Classification**

As we know some objects share similar properties, so based on their properties, it is checked with the existing objects property. If it matches it is added into its group, otherwise neglected. This process is called classification. Data classification consists of two-steps, one is learning step and another is classification step. The first step consists of a classifier in which existing object property is stored. In next step, the algorithm scrutinizes the object with the existing object properties and classifies its category. A tuple Y is depicted by a x-dimensional attribute vector, Y= (y1, y2 ,… , yn.). As we know every tuple Y belongs to a pre-specified class. Since we are already know its class table in advance, so it is also called as supervised.

Now the computation is done for forecasting the correctness of a classifier. The test set do not depend on the training tuples, it means that they do not play any role in constructing the classifier. An instance of a classification task is classifying customers based on their monthly salary. When the customers in a bank apply for a loan they are asked several questions for example salary, duration of employment, present address. These queries answers help the loan manager to decide whether the applicant should be funded or not. This acts like a classification method. Data mining techniques which help in classifications are decision trees and nearest neighbor techniques.

- **Estimation**

A A shopkeeper keeps the record of a sell of a particular item. He feels that the particular item is sold 2 pieces per day. It means that he will keep in stock at least 60 pieces in a month. The prediction of total value for a product for a specific period of time is called estimation. Regression and neural networks are the techniques which can be applied for estimation in data mining.

- **Association Extraction and Frequent pattern**

Association extraction is a task which generates the items which occur together depending upon the occurrence of an item. The market basket analysis is one of the best instances of association extraction between the item set. For instance, suppose a customer wants to a bucket at the general store, there could be possibility to purchase a mug with it? If the shopkeeper knows what items a customer wants to purchase together can help general stores with attractive advertising, storage in the shop, pricing, promotions and inventory management. Frequent pattern is a method which is related to association extraction between the items but the objective is to identify the most frequent items together over a list of transaction. An antecedent is an event which occurs always before the consequent .The pattern can be symbolized as:
Consequent ⇐ Antecedent
is up to the analyzer that how many antecedent they want to use in a sequence. The most popular and demanding data mining technique for extraction between items and frequent pattern is the CARMA algorithm.

- **Clustering**

Clustering is a method of splitting a group of data objects into subsets. Every subset builds a cluster, in such an order that cluster objects are similar to one another. As clustering breaks large item set into smaller groups, it is also called as data segmentation. It can also be applied for outlier detection in handwritten character recognition systems for image recognition. For example the two scientists evaluated the star's temperature and its brightness. They found the three different groups of stars. Each group has totally different feature from the others. These helped them to cluster the stars having similar attributes.

1. Problem statement: "Interpolation of Stock Market Data with Fuzzy Conception Using Weka Tool" 2.3.2 Objectives:

1. The proposed algorithm calculates variance and standard deviation which enables us to decide that by how much the stock prices units would fluctuate about the mean stock price over a period of time
2. It allows the investors to decide the selling and buying of products by watching the fluctuation of stock prices over a period of time
3. Prediction of risk associated with products.

A. 4.2 Problem Formulation and Proposed Solution

Input: A time series TS with n data points, a list of m membership functions for data points, a predefined minimum support threshold α, a predefined minimum confidence threshold λ, and a sliding window size ws .

STEP 1: Calculate the mean of given time series TS with n data points.

$$\text{Mean } (\bar{x}) = 1/\text{n} \sum_{i=1}^{n} x_i$$

STEP 2: Determine the variance of n data points of time series TS

$$\text{Variance } (\sigma^2) = \sqrt{\sum_{i=1}^{n} 1/n(x_i - \bar{x})^2}$$

STEP 3: Calculate the standard deviation of K data points

Standard deviation

$$(\sigma) = 1/n \sum_{i=1}^{n}(x_i - \bar{x})^2$$

STEP 4: Convert the time series TS into a list of subsequences W (TS) according to the sliding-window size ws. That is, W (TS) = {sb | sb = (db, db+1,……, db+ws-1), b = 1 to (n−ws+1),where db is the value of the b-th data point in TS.

STEP 5: Transform the k-th (k = 1 to ws) quantitative value vbk in each subsequence sb (b = 1 to n−ws + 1) into a fuzzy set fbk represented as (fbk1/Rk1 + fbk2/Rk2 +. . . + fbkn/Rkn) using the given membership functions, where Rkl is the l-th fuzzy region of the k-th data point in each subsequence, m is the number of fuzzy memberships, and fbkl is *v*bk's fuzzy membership value in region Rkl. Each Rkl is called a fuzzy item.

STEP 6: Compute the scalar cardinality of each fuzzy item Rkl as

$$\text{Count}_{\text{kl}} = \sum_{b=1}^{\text{n} -\text{ws}+1} f_{bk1}$$

STEP 7: Group the above obtained fuzzy items to form the candidate 1-itemsets C1.

STEP 8: Check whether the support value (=count$_{k1}$/n−ws + 1) of each Rkl in C1 is greater than or equal to the predefined minimum support threshold α. If Rkl fulfil the above condition, collect it in the set of large 1-itemsets (L1). That is:
L1 = {Rkl |count$_{k1}$ ≥α,1≤k≤b+ws−1 and 1≤l≤m}.

STEP 9: IF L1 is not null, then perform the next step; otherwise, terminate the algorithm.

STEP 10: Set t = 1, where t is used to represent the number of fuzzy items in the current item sets to be processed.

STEP 11: Join the large t-itemsets Lt to obtain the candidate (t+ 1)-itemsets Ct+1 in the same way as in the Apriori algorithm provided that two items obtained from the same order of data points in subsequences cannot exist in an itemset in Ct+1 at the same instant.

STEP 12: Proceed the following substeps for each newly formed (t + 1) - itemset I with fuzzy items (I1, I2, . . ., It+1) in Ct+1:
(a) Compute the fuzzy value of I in each subsequence sb as $f_I^{sb} = f_{I1}^{sb} \wedge f_{I2}^{sb} \wedge f_{I3}^{sb} \wedge \ldots\ldots \wedge f_{It+1}^{sb}$ is the membership value of fuzzy item Ik in Sb . If the minimum operator is used for the intersection, then:

$$f_I^{sb} = \text{Min}_{k=1}^{t+1} f_I^{sb}$$

If the support (=count$_\text{I}$/n−ws + 1) of I is greater than or equal to the predefined minimum support threshold α, put it in Lt+1.

STEP 13: If Lt+1 are null, then do the next step; otherwise, set t= t + 1 and repeat STEPs 11–

STEP 14: Generate the association rules for each large h-itemset I with items (I1, I2, . . .,Ih),
h≥2, using the following sub steps:

a) Form each possible association rule as follows:

$$I_1 \wedge \ldots \wedge I_{n-1} \wedge I_{n+1} \wedge \ldots \wedge I_h \rightarrow I_n,$$

n = 1 to h.

(b) Calculate the confidence values of all association rules by the following formula:

$$\sum_{b=1}^{n-ws+1} f_I^{sb} / \sum_{b}^{\text{n}-\text{ws}+1} (f_I^{sb} \wedge \ldots . \wedge f_I^{sp})$$

STEP 15: (a) Output the fuzzy association rules with confidence values greater than or equal to the predefined confidence threshold λ. from time series data points TS.

(b) Measurement of distribution of data points along their mean provided the mean is chosen as the center point.

B. Comparison Results

The proposed work has shown comparison with the base reference work [1].

CASEI- We have plotted data pixel according to the dataset of stock market .A bigger dataset has been chosen to increase the efficiency of the proposed algorithm. The fig below shown the graph of scattered data points for minimum confidence vs. membership function, unlike the previous work the classification is done more accurately and dynamically comparatively.

CASE II- The next graph has shown different content thoroughly in which there is less ambiguity than the previous work as the Graph has shown three different class showing with different colour clearly. So this dynamically allotted class is more pertinently working with clear visualization.

CASE III-This case focuses on randomly allocated stock market data set which are inconsistent in nature and one can’t easily perform classification and predication. In the previous work the dataset was too small to explore the dynamic dimension of any operation of Algorithm.

CASE IV –The resultant of comparison should be efficient enough to stand by some specific result .On a dynamic dataset of stock market with variable sliding window, our proposed work has applied efficiently and finally shown in the figure. The plot has shown exactly n region with cluster result on a dynamic set unlike the ambiguous result of previous algorithm. The strength of this proposed technique is less ambiguity and more accuracy.

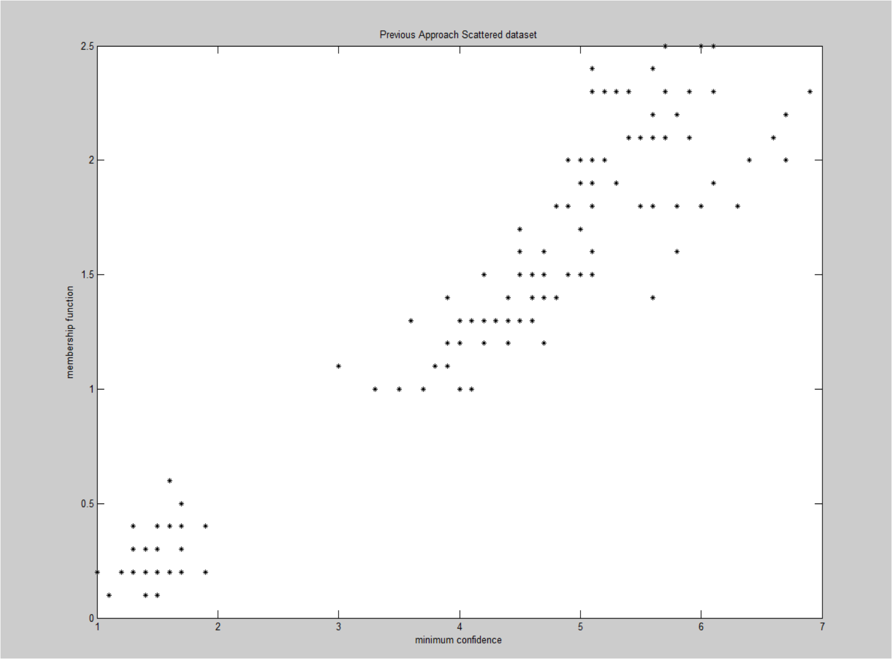

Fig: Scattered Data Points for Minimum Confidence vs. Membership Function

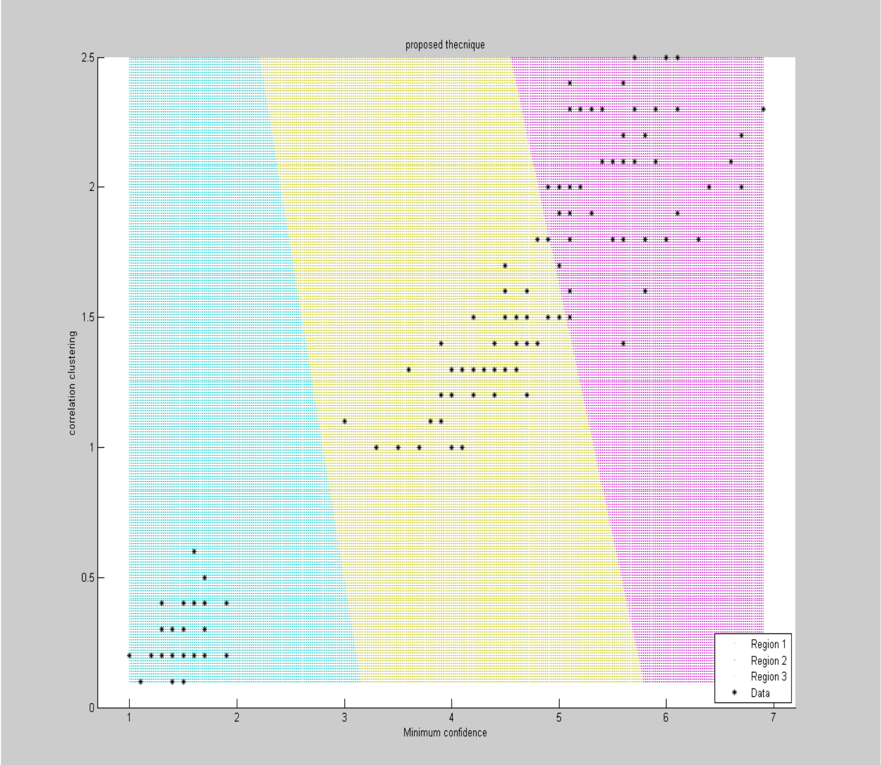

Fig: Graph has shown three different classes

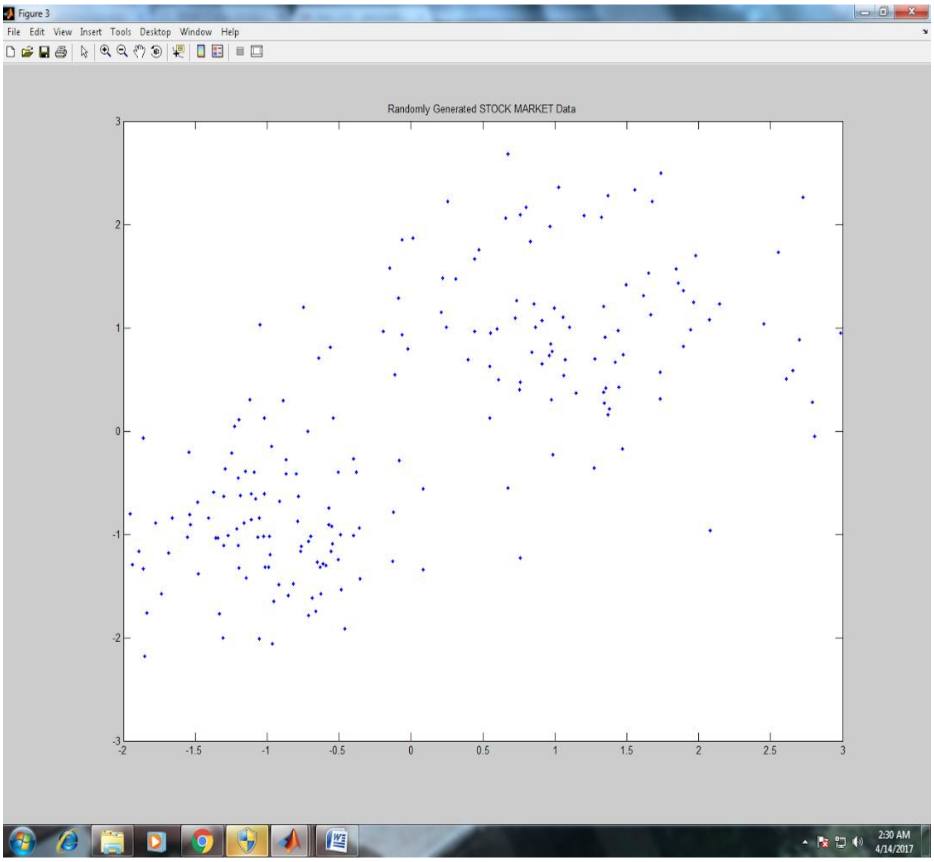
Figure 3
File Edit View Insert Tools Desktop Window Help
Randomly Generated STOCK MARKET Data
3
2
1
0
-1
-2
-3
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
2:30 AM
4/14/2017

Fig: Randomly Allocated Stock Market Data Set

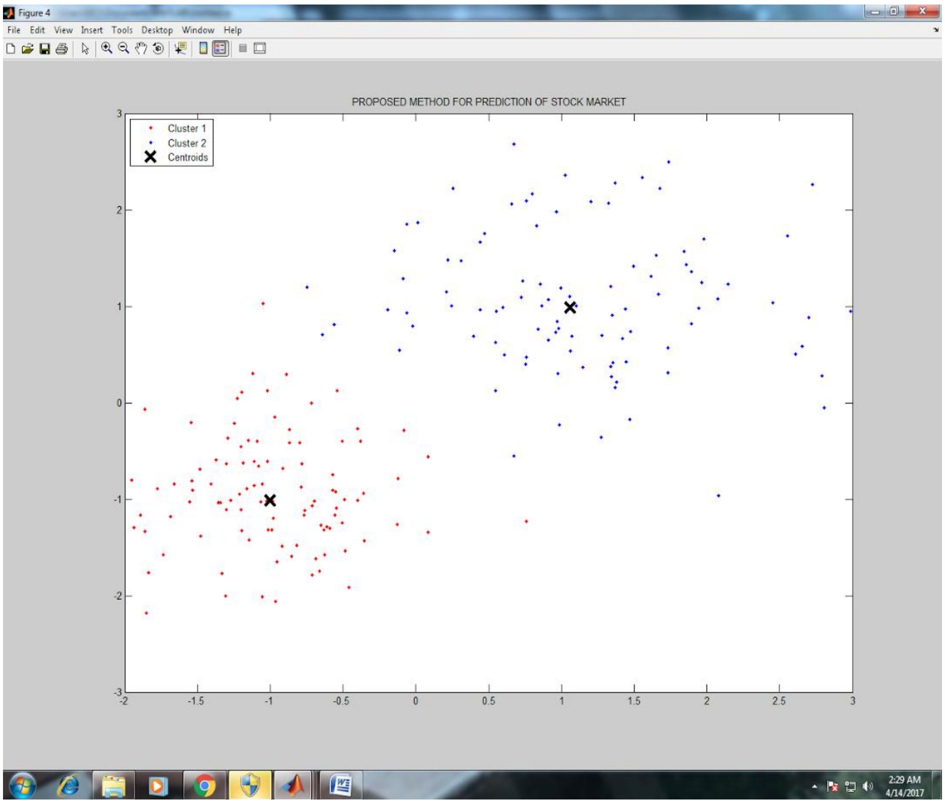
Figure 4
File Edit View Insert Tools Desktop Window Help
PROPOSED METHOD FOR PREDICTION OF STOCK MARKET
Cluster 1
Cluster 2
Centroids
3
2
1
0
-1
-2
-3
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
2:29 AM
4/14/2017

Fig: Dynamic Dataset of Stock Market with Variable Sliding Window

II. RESULTS AND CONCLUSION

Throughout the paper it is attempted to induce the fuzzy association rules and reduce the irrelevant fuzzy rules. Apart from this, the proposed algorithm showed the stock price dispersion from the mean stock price over a period of time which would help the investor to understand the market fluctuation. It would also explain at what point of time the selling and buying activity could be done. The proposed approach can remove lots of redundant rules through proper filtering process, so that users can effectively access the rules. Future work suggests that the membership function can be set dynamically. In this paper membership functions are known in advance. More complex operations could be made in near future.

REFRENCES

[1]. Jiawei Han, Jian Pei, ―Data Mining Concepts and Techniques‖, ISBN-978-93-80931- 91 3, Elsevier 2013

[2]. Abdullah Al Mueen, ―Exact Primitives for Time Series Data Mining‖ University of California riverside 2014

[3]. B.Liu, W.Hsu, YMa, ―Mining Association Rules with Multiple Minimum Supports‖, International Conference on Knowledge Discovery and Data Mining, pp-337-341 1999

[4]. R.Agrawal, R.Srikant,‖Fast Algorithm for Mining Association Rules‖, International Conference on Very large Databases, pp - 487-499, 1994

[5]. R.Agrawal, R.Srikant, ―Mining Sequential Patterns‖, The 11th International Conference On Data Engineering, pp-3-14, 1995

[6]. J-S.R.Jang, C-T.Sun, ―Neuro-Fuzzy and Soft Computing‖, ISBN-978-81-203-2243-1, PHI, 2011

[7]. T P Hong, K Y Lin and S L Wang, ―Fuzzy Data Mining for Interesting Generalized Association Rules‖, Fuzzy Sets & Symbols, Elsevier pp-255-269 2002

[8]. Cai, ―Mining Association Rules with Weighted Items‖ International Database Engineering and Applications Symposium 1998

[9]. T P Hong, K Y Lin, ―Induction of Fuzzy Rules and membership Functions from Training Examples‖, Fuzzy Sets and Systems, Elsevier pp-33-47, 1996

[10]. Richard J.Povinelli, ―Time Series Data Mining: Identifying Temporal patterns for Characterization and prediction of Time series and Events‖ Doctoral Thesis, Marquette University 1999

[11]. Christopher J. Neely, ―Technical Analysis in the Foreign Exchange Market: A Layman's Guide‖ A Review, 1997

[12]. R.Srikant, R.Agrawal, ―Mining Quantitative Association Rules in Large Relational Tables‖, International Conference on Management of Data, ACM pp-1-12, 199 47

[13]. Au, Chan, ―Mining Fuzzy Rules for Time Series Classification‖ International conference on Fuzzy system 2004 G. Das, K Lin, ―Rule Discovery from Time Series,‖ in: Proceedings of the 4th International Conference on knowledge discovery and data mining pp-16-22 1998