IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 8, Issue 5 (Jan. - Feb. 2013), PP 22-27
www.iosrjournals.org

Analysis of Time Series Rule Extraction Techniques

Hima Suresh [1], Dr. Kumudha Raimond [2]
[1](PG Scholar/Dept. of CSE, Karunya University, India)
[2](Professor/Dept. of CSE, Karunya University, India)

Abstract : *In Data mining, the sequence of data points are measured typically at successive time instants spaced at uniform time intervals are called time series. The real applications of time series are frequent pattern analysis, bioinformatics, medical treatment, meteorology, sociology and economics. Frequent patterns can be analyzed to give explanatory rules and this rule extraction can be done using many algorithms like Genetic Algorithm, Fuzzy Logic , Support Vector Machine etc. Rule induction is an area of machine learning in which rules are extracted from collective set of observations. The rules extracted may represent complete scientific model of the data, or simply represent local patterns in the data. A brief overview of some of the most common rule extraction techniques and a comparison between single and hybrid rule approaches comprise in this survey.*

Keywords- *Discrete-Wavelet Transform (DWT), Fuzzy Logic (FL), Genetic Algorithm (GA), Neural Network (NN), Support Vector Machine (SVM).*

I. INTRODUCTION

Time Series is a collection of observations of well defined data items which are obtained through repeated measurements over time intervals. It has got lot of applications such as analysis of frequent patterns, predictions etc. Analysis of frequent patterns in time series data has become one of the core data mining tasks and has attracted tremendous interest among researchers. Frequent patterns can be analyzed to give explanatory rules which can be done using many algorithms. An algorithm which integrates the fuzzy sets, the apriori algorithm, and the time series concepts to find out appropriate linguistic rules proposed for analysis of time series by C.H.Chen.et al [1]. There have been many approaches to reduce the dimensionality while preserving the most important "features" in the reduced basis set. Uses of DWT (Discrete-Wavelet Transform) are proposed by J.Schott.et al [2]. Neural Network (NN) methods for time series analysis are proposed by Z. Zhang [3]. This technical report provides a survey on the existing methods to analyze time series rule extraction methods.

The remainder of this paper is organized as follow: Section II presents the rule extraction using various techniques. Section III presents Experimental analysis. Finally, conclusion is discussed in Section IV.

II. RULE EXTRACTION TECHNIQUES

Rule Extraction plays an important role in data mining. There are several rule extraction techniques. The Rules are extracted for the purpose of time series analysis. The existing methods used either single or hybrid approaches.

1. Single Approaches

Single approaches are using simple algorithms for extracting rules. An approach was proposed by G.N.Pradhan.et al [4] which considering real-life time series data of muscular activities of human participants which are obtained from multiple electro myogram sensors (EMG) and discovers patterns in these EMG data streams. A two stage approach is proposed for this purpose. First stage emphasizes discovering patterns in multiple time series by doing sequential mining across time slices. In the next stage, it focuses on the quantitative attributes of only those time series that are present in the patterns discovered in the first stage. The method speeds up the process of finding association rules .This approach is generic and can be applicable to any multiple time series dataset format.

A Genetic algorithm based approach was proposed by B.M. Al-Maqaleh [5] for mining classification rules from large data base and those rules discovered by the algorithm have higher classification performance to unknown data.

A Classification algorithm based on Evolutionary Algorithm (EA) which could discover possible interesting small –disjunct rules in the form If P then D was proposed by B.M. Al-Maqaleh.et al [6]. The proposed algorithm is validated on several data set repository and the experimental results are presented to demonstrate the effectiveness of the proposed scheme for automated small – disjunct rules mining. The EAs has

been applied to Knowledge Discovery in Databases with the motivation that they are robust and adaptive .These search methods performs a global search in the space of candidate solutions.

An association rule mining technique that have been recently developed and used for genomic data analysis proposed by M. Ananthavalli et.al [7]. Global Gene Expression data can be a valuable tool in understanding of biological networks, genes and cellular states. Two important goals are there for the analysis of these massive genomic data: First goal is to determine how the particular gene might affect the expression of other genes. Genes involved in this case could possibly belong to the same gene network. Second goal is to determine the possible kind of genes that are expressed as a result of cellular conditions.

A single Fuzzy approach for extracting general temporal association rules in a publication database has been proposed by G.C. Lan et.al [8]. The algorithm describes the life span of an item which is measured by its entire publication periods in a publication database. To keep and obtain information of itemsets effectively for data mining, an itemset table structure has been designed. The proposed algorithm finds more frequent itemsets from the given dataset.

An Efficient Distributed Genetic Algorithm for classification rule mining has been proposed by M. Rodriguez et.al [9]. This promotes a new method of distribution of data in computer networks. The process is actually done by spatial partitioning of the population into different semi-isolated nodes, each of the nodes involving in parallel and possibly exploring several regions of the search space. Two techniques were proposed to generate an accurate classifier with data partition: an elitist pool which is for the selection of rule and a novel technique of distribution of data (DLF). To dynamically redistribute the training data in the node neighborhood, DLF uses heuristics based on the local data. This result shows a considerable speed up and this improvement which does not compromise the classifier's accuracy and complexity.

2. Hybrid Approaches

Hybrid approaches are the combination of different algorithms. It is used to overcome the limitations of single approaches. C.H. Chen [1] proposed a fuzzy mining algorithm which integrates the fuzzy sets, the apriori algorithm, and the time-series concepts to find out appropriate linguistic association rules. The proposed approach has two advantages. Firstly they will be friendlier to human than quantitative representation, since the final results are represented by linguistic rules. Secondly, one problem for association rule mining approaches is that too many rules may be generated. Lots of redundant rules can be filtered through the post-processing in the algorithm such that the rules can be compact. A Rule Extraction approach to obtain maximum accuracy for the prediction of diabetes was proposed by G.Suganya.et al [10]. Support Vector Machines (SVM) is utilized for the diagnosis of diabetes. An additional explanation module used which turns the black box model of an SVM into an intelligible representation of the SVM's diagnostic decision. Results on a real-life diabetes dataset show that intelligible SVMs provide a promising tool for the prediction of diabetes, where a comprehensible rule set have been generated . Here two techniques are used for rule extraction. Sequential covering Rule Extraction and the eclectic methods are used to turn the SVM black box into a more intelligible model. The hybrid system for medical diagnosis was developed.

A framework proposed by J.Schott.et al [2] that utilizes an Adaptive Network based Fuzzy Inference System (ANFIS) to perform user constrained pattern recognition on time series data. The architecture used here allows an analyst to represent domain expertise in a context relevant manner. Fuzzy logic rules constructed are used to perform feature extraction and it influences the training of a Neural Network (NN) in order to perform pattern recognition. This architecture is also capable of performing noise tolerant searches across multiple features on large volumes of time series data.

An efficient neuro–fuzzy-genetic data mining frame work based on computer intelligence was proposed by Z. Zang [3], which combines computational intelligent tools such as NN, Fuzzy Logic and GA .This framework discovers patterns and represents them in understandable forms. In order to extract explicit knowledge from the trained NNs and represent it in the form of fuzzy if-then rules, a rule extraction technique is applied. In the final stage, a GA is used as a rule –pruning module to eliminate those weak rules that are still in the rule bases.

An algorithm proposed by A.M. Palacios et.al [10] which integrates imprecise data concepts and the fuzzy apriori mining algorithm to find interesting fuzzy association rules in given data bases. This is proposed with the aim of getting high quality fuzzy association rules from databases with interval and fuzzy values. To do this, several important aspects have been considered due to the true value of one data being unknown and the fuzzy membership value interpreted as a set of bounds of probabilities. These rules from low quality provide knowledge about the dependencies and relationship between the items and therefore several items can be excluded or removed due to their being considered irrelevant. Single and Hybrid approaches have their own limitations. By using the combination of different algorithms, would give better performance that comprise hybrid approach.

III. EXPERIMENTAL ANALYSIS

This section details the results of simple apriori algorithm using a time series data set and the comparison of hybrid approaches with an integration of both fuzzy and apriori algorithm. The time series data set used for implementing apriori algorithm is shown in the Table 1:

Table 1: the time series data

Years	Home price	Years	Home price
1999	127	2006	124
2000	129	2007	118
2001	132	2008	121
2002	130	2009	120
2003	126	2010	115
2004	132	2011	113
2005	129	2012	119

According to the window size different subsequences of transactions are obtained. Here window size is assumed as 5. The transaction of the time series data is shown in Table 2:

Table 2: Transactions of time series data

{127,129,132,130,126}	{132,129,124,118,121}
{129,132,130,126,132}	{129,124,118,121,120}
{132,130,126,132,129}	{124,118,121,120,115}
{130,126,132,129,124}	{118,121,120,115,113}
{126,132,129,124,118}	{121,120,115,113,119}

The frequent items are generated from this and their support values are calculated as proposed by Agrawal et.al [12]. The items which are frequently obtained and their support values are shown in the Table 3:

Table 3: The 1-item set and their support values

Item	Support	Item	Support	Item	Support
{129}	7/10	{124}	5/10	{115}	3/10
{132}	6/10	{118}	5/10	{113}	2/10
{130}	4/10	{121}	5/10	-	-
{126}	5/10	{120}	4/10	-	-

Assuming the minimum support as 5/10, So that the 2-item set to be generated which satisfies the support value. The 2-item set and their corresponding support values are shown in the Table 4:

Table 4: The 2-itemsets and their supports

Item	Support	Item	Support	Item	Support
{129,132}	6/10	{132,115}	0/10	{124,118}	4/10
{129,130}	4/10	{132,113}	0/10	{124,121}	3/10
{129,126}	5/10	{130,126}	4/10	{124,120}	2/10
{129,124}	4/10	{130,124}	1/10	{124,115}	0/10
{129,118}	3/10	{130,118}	0/10	{124,113}	0/10
{129,121}	2/10	{130,121}	0/10	{118,121}	4/10
{129,120}	1/10	{130,120}	0/10	{118,120}	3/10
{129,115}	0/10	{130,115}	0/10	{118,115}	2/10
{129,113}	0/10	{130,113}	0/10	{118,113}	1/10
{132,130}	4/10	{126,124}	2/10	{121,120}	4/10
{132,126}	5/10	{126,118}	1/10	{121,115}	3/10
{132,124}	3/10	{126,121}	0/10	{121,113}	2/10
{132,118}	2/10	{126,120}	0/10	{120,115}	3/10
{132,121}	1/10	{126,115}	0/10	{120,113}	2/10
{132,120}	0/10	{126,113}	0/10	{115,113}	2/10

The pairs that meet or exceed the minimum support of 5/11 are shown in the Table 5 below.

Table 5: The 2-itemsets with support ≥ 5/11

Item	Support	Item	Support	Item	Support
{129,132}	6/10	{129,126}	5/10	{132,126}	5/10

The 3-item set obtained from the corresponding 2-item sets are shown in the Table 6:

Table 6: The 3-item sets and their support value

Item	Support
{129,132,126}	5/10

Therefore the frequent items obtained are {129},{132},{130},{126},{124},{118},{121}, {120},{120}, {115}, {113},{129,132},{129,126},{132,126},{129,132,126}. Finally rules are generated and their corresponding confidence values are calculated using the formula:

$$Confidence = \frac{Support(XUY)}{Support(X)} \quad (1)$$

Let the confidence be >85%. The Rules which meet or exceed the corresponding confidence has been taken and is shown in the Table 7 below:

Table 7: Rules generated using apriori

Rules	Support (XY)	Support (X)	Confidence
{126}-> {132}	5/10	5/10	100
{126}-> {129}	5/10	5/10	100
{132}-> {129}	6/10	6/10	100
{126}->{132,129}	5/10	5/10	100

This is an example of single approach which is using apriori algorithm to generate rules. Here the rules are specific and are temporal. First rules specifies that if {126} occurs then {132} occurs. Likewise if {126} occurs then {129} and so on. This method doesn't specify the time units from which the rule occurs. The method provides only general rules.

The Hybrid approach using both fuzzy and apriori algorithm is an efficient method for handling time series data to find linguistic association rules [1][11]. The time series data used here is the same Real homeprice data over years from 1999 to 2012 in order to compare the performance with respect to single approach and are shown in the Table 8:

Table 8: Time series data

Years	Home price	Years	Homeprice
1999	127	2006	124
2000	129	2007	118
2001	132	2008	121
2002	130	2009	120
2003	126	2010	115
2004	132	2011	113
2005	129	2012	119

The time series data is transformed into subsequences according to the window size which is predefined. Here in this example window size is taken as 5. So the total subsequence will be 10 according to the formula (total time series – window size+1). The subsequence generated is shown below in Table 9:

Table 9: Subsequences generated from time series

S	Subsequence	S	Subsequence
S1	(127,129,132,130,126)	S6	(132,129,124,118,121)
S2	(129,132,130,126,132)	S7	(129,124,118,121,120)
S3	(132,130,126,132,129)	S8	(124,118,121,120,115)
S4	(130,126,132,129,124)	S9	(118,121,120,115,113)
S5	(126,132,129,124,118)	S10	(121,120,115,113,119)

After transforming time series data into subsequence, the data values in each sub sequences are converted to fuzzy sets according to predefined membership function. The fuzzy set transformed is shown in the Table 10 below:

Table 10: Fuzzy set generated for time series data

S	A1L	A1M	A1H	A2L	A2M	A2H	A3L	A3M	A3H	A4L	A4M	A4H	A5L	A5M	A5H
S1	0	1	0	0	0.67	0.33	0	0	1	0	0	0	0	1	0
S2	0	0.67	0.33	0	0	1	0	0	0	0	1	0	0	0	1
S3	0	0	1	0	0	0	0	1	0	0	0	1	0	0.67	0.33
S4	0	0	0	0	1	0	0	0	1	0	0.67	0.33	0	1	0
S5	0	1	0	0	0	1	0	0.67	0.33	0	1	0	1	0	0
S6	0	0	1	0	0.67	0.33	0	1	0	1	0	0	0.67	0.33	0
S7	0	0.67	0.33	0	1	0	1	0	0	0.67	0.33	0	0.67	0.33	0
S8	0	1	0	1	0	0	0.67	0.33	0	0.67	0.33	0	1	0	0
S9	1	0	0	0.67	0.33	0	0.67	0.33	0	1	0	0	1	0	0
S10	0.67	0.33	0	0.67	0.33	0	1	0	0	1	0	0	0.67	0.33	0
total	1.67	4.67	2.66	2.34	4	2.66	3.34	3.33	2.33	4.34	3.33	1.33	5.01	3.66	1.33

The total count or scalar cardinality has been calculated as its count value. After that the fuzzy items generated and are collected as the 1-item sets. The total count is then checked against the support value, which is predefined. Here in this example support value is set as 34%. The itemsets whose values are greater than or equal to 34% has been taken and are the following: A1.Middle, A2.Middle, A4.Low, A5.Low, and A5.Middle. This is taken as the set of large 1-itemsets. The itemset is shown in the Table 11 below:

Table 11: Generation of 1-itemset with total count

1-Itemset	Total	1-Itemset	Total	1-Itemset	Total
A1.M	0.467	A2.M	0.4	A4.L	0.434
A5.L	0.501	A5.M	0.366	-	-

After this 2-itemsets are generated from the corresponding 1-itemsets.The fuzzy items with the same attribute Ai are not put in to 2-itemset. 2-itemset generated is shown in the Table 12 below:

Table 12: Generation of 2-itemsets and their counts

Itemset	Total	Itemset	Total	Itemset	Total
A1.M∩A2.M	1.67	A2.M∩A5.L	2	A2.M∩A4.L	2
A1.M∩A4.L	1.67	A2.M∩A5.M	2.66	A4.L∩A5.L	3.68
A1.M∩A5.L	3	A4.L∩A5.L	3.68	-	-
A1.M∩A5.M	1.66	A4.L∩A5.M	0.99	-	-

The association rules are then generated from the 2-itemsets.The association rule generated is:

1) **If A4= Low, then A5= Low with confidence 0.368; the** confidence value of above rule is calculated using the formula given below:

$$\frac{\sum_{P=1}^{11}(A_1 .Middle \cap A_4 .Middle)}{\sum_{P=1}^{11} A_1 .Middle)} \quad (2)$$

The confidence value of rule is then compared with the predefined confidence threshold. The rule obtained here means that "if the value of a data point is middle then the value of a data point after three time units will also be middle with a high probability".

Comparing apriori and fuzzy apriori method, fuzzy apriori method is more efficient. Fuzzy apriori method gives general rules. It clearly specifies the time units. Whereas apriori method gives specific rules and is a temporal rule extraction method but it does not specify the time units.

IV. Conclusion

The paper discusses about various rule extraction techniques. The main aim of this inspection is to explore highly efficient method for generating non redundant rules. This technical report provides rule mining on the existing methods based on single approaches and hybrid approaches. Each rule extraction techniques have advantages and disadvantages. This survey report shows that some existing methods consider only single

approaches while other methods consider the combination of single approaches. Mostly GA, Fuzzy Apriori ,NN etc are used for extracting rules which can be considered as single approaches and the combination of such approaches contribute hybrid ones. Finally examples of single and hybrid methods are compared and concluded that the hybrid approaches have high performance than single approaches.

REFERENCES

[1] C. Chen, T. Hong, V.S. Tseng, "Fuzzy datamining for time series data", *Soft Computing* - Vol.12,pp , 536-542, 2012.

[2] J. Schott, J. Kalita, "Neuro fuzzy time series analysis of large volume data", *Intelligent Systems in Accounting, Finance and Management*, Vol. 18, pp 39–57, January/March *2011*.

[3] Z.Zhang, "An efficient neuro-fuzzy-genetic data mining framework based on computational intelligence", Vol. 2, pp 178-183, Aug 2009.

[4] G.N. Pradhan, B. Prabhakaran, "Association rule mining in multiple, multidimensional time series medical data ", *IEEE International Conference on Multimedia and Expo*, pp 1716- 1719, Dec 2009.

[5] B.M.A.Maqaleh, H. Shahbazkia ," A GA for discovering classification rules in data mining *International Journal of Computer Applications* , Vol. 41, March (2012).

[6] B.M. Al-Maqaleh, M.A. Al-Dohbai, H. Shahbazkia, "An Evolutionary Algorithm for automated discovery small disjunct rules", *International Journal of Computer Applications*, Vol.41 , March 2012.

[7] M. Anandhavalli, M.K. Ghose, K. Gauthaman, "Association rule mining in genomics", *International Journal of Computer Theory and Engineering*, Vol.2, pp 1793 -8201, April 2010.

[8] G.C. Lan, C.H. Chen, T.P. Hong, S.B. Lin, "A fuzzy approach for mining general temporal association rules in a publication database", *International Conference on Hybrid Intelligent Systems*, 2011.

[9] M. Rodriguez, D.M. Escalante, A. Peregrin, "Efficient distributed Genetic Algorithm for rule extraction ", *Applied soft computing*, Vol. 11, pp 733 – 743, January 2011

[10] G. Suganya, D. Dhivya, "Extracting diagnostic rules from support vector machine", *Journal of Computer Applications*, Vol.4, 2011.

[11] A.M. Palacious, M.J. Gacto, J.Alcala-Fdez, "Mining fuzzy association rules from low quality data ", *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, Vol.16, pp 883-901, 2012.

[12] Agrawal, R., Imielinski, T., Swami, A,: Mining association rules between sets of items in large databases. In: SIGMOD, pp.207 - 216.Washington D.C.(USA)(1993)