Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 7, July -2017

Functional Association Rule Mining Using Cooperative Coevolutionary Deep Neural Networks

Ms. Pooja Kulkarni[1], Mrs. Mrs. V. L. Kolhe[2]

[1]Department of Computer Engineering, D. Y. Patil College of Engineering, Akurdi
[2]Department of Computer Engineering, D. Y. Patil College of Engineering, Akurdi

Abstract — *Functional Association rule form is novel form of association rules (ARs) that do not require discretization of continuous variables or the use of interval values in either sides of the rule. This rule form captures nonlinear relationships among continuous variables, and provides an alternative pattern representation for mining essential relations hidden in a given data. A new neural network based, co-operative, coevolutionary algorithm is presented for FAR mining. Conventionally, ARM is majorly concerned with categorical data sets. When it is used to process continuous variables, it converts the values of the variables into intervals. This discretization process determines the granularity of the ARs being generated and generates granularity levels for ARs. In contrast, the FAR proposed can handle nonlinearity in the relationship and can deal with continuous variables directly and without converting them into intervals. Deep Learning approach is used to increase the accuracy of FAR mining. A new measure for accuracy is introduced. K-means clustering used for normalization of the continuous data.*

Index Terms—*Association rules mining (ARM), co-evolutionary algorithms (EAs), neural networks, predictive models.*

I. INTRODUCTION

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration. Data mining can be viewed as a result of the natural evolution of information technology. The database system industry has witnessed an evolutionary path in the development of the following functionalities: data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and advanced data analysis (involving data warehousing and data mining). For instance, the early development of data collection and database creation mechanisms served as a prerequisite for later development of effective mechanisms for data storage and retrieval, and query and transaction processing. With numerous database systems offering query and transaction processing as common practice, advanced data analysis has naturally become the next target [4]. Mining, a shorter term, may not react the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer that carries both data and mining became a popular choice. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging [4].

Association rules are if/then statements that help find relationships between seemingly unrelated data in a relational database or other information repository. One of the reasons behind maintaining any database is to enable the user to find interesting patterns, information and trends in the data. For example, in a supermarket, the user can figure out which items are being sold most frequently. But this is not the only type of 'trend' which one can think of. The goal of database mining is to automate this process of finding interesting patterns and trends. The output of the data mining process should be a summary of the database [5].

Programmers use association rules to build programs which are capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient and accomplished with-out being explicitly programmed. The current expert system technologies, which typically rely on users or domain experts to manually input knowledge into knowledge bases. This procedure contains errors, and it is extremely time consuming and costly. Data mining tools which perform data analysis may find important data patterns. Data mining
represents the automatic process to discover patterns and relations between data stored in large databases called warehouses, the final product of this process being the knowledge, meaning the significant information provided by the unknown elements AR mining (ARM) is an important branch of methods for extracting patterns from data sets. Conventionally, ARM is concerned with categorical data sets. When it is used to process continuous variables, it converts the values of the variables into interval values. This discretization process determines the granularity of the ARs being generated. In contrast, the proposed FAR can handle nonlinearity in the relationship and can deal with continuous variables directly and without converting them into intervals. Therefore, it does not require a discretization process,

neither does it need predetermining granularity levels. A co-operative, co-evolutionary algorithm (EA)-based mining scheme for mining FARs is proposed. First, FAR captures the associative relations among the variables of a given data set rather than the intervals of the variable values. Second, the key representation component of FAR is based on artificial neural network (ANN), which is the foundation for FAR to capture quantitative associative nonlinear relations hidden in the data set. ANNs have played a central role in classification and prediction problems. It provides a robust representation to identify real valued hidden associative relations. Third, due to the predictive power of ANN, FAR can increase the granularity of ARs being generated, as it directly predicts the values of its right-hand side (RHS) variables. Fourth, the mining algorithm for FAR is inherently parallel [1].

The section II gives review of literature. In section III, system architecture is explained. Section IV gives system analysis. Results are given in section V.

II. REVIEW OF LITERATURE

Extracting patterns from raw data assists analysts to under-stand the underlying regularities hidden in the variations in the data. Such patterns are usually represented by different basic models, including classifiers, regression functions, and rules. Among them, ARs have received extensive investigations during the past two decades due to their simplicity, com-prehensibility, accuracy and close connection with the real world applications. ARM was first introduced in [1], and is typically used together with transactional data sets. A frequent pattern is found interesting when the number of transactions containing it exceeds a predefined minimum threshold called the minimum support.

III.

Research on mining ARs is done to investigate various properties of associative rules, including infrequent rules [14], weighted rules [15], and sensitive information [16], and different applications, including anomaly detection [17], bioinformatics [18], and text mining [19].

IV.

In the literature, the former is more interesting for re-searchers , due to its compatibility with classical ARM approaches. In this branch, some methods use discretization as a preprocessing step, where it is usual to include domain knowledge, binning, and clustering.

V.

Optimization-based partition techniques are then developed. Mata et al. [4] proposed a genetic algorithm (GA)-based method, genetic AR (GAR), to optimize the support of frequent patterns directly on raw variable values. The intervals of attributes and members of frequent patterns are dynamically adjusted during evolution. GARs optimization focuses on frequent patterns, which does not take care of confidence status of derived rules. QuantMiner [20] extended this study for optimization to be carried out directly on ARs incorporating confidence measures into its objective function. Rules with high support and high confidence are favored during evolution. However, it needs a specification of rule templates for the mining process. Alatas et al. [5] proposed the multiobjective differential EA for mining numeric ARs (MODENAR) to mine ARs for continuous variables. Each individual encodes a potential rule without specifying any special form of rules. Here, four objectives (support, confidence, rule comprehensibility, and interval amplitude) are used to identify the optimal rules. Genetic programming for mining continuous variable is used in [21]. Fuzzy sets are also used as a tool to overcome the problems brought by discretization [22], [23].

III. SYSTEM ARCHITECTURE / SYSTEM OVERVIEW

3.1 Rule Generation

Apriori Algorithm: Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets appear suciently often in the database. The frequent itemsets determined by apriori are used to determine functional Association rules.

Indices are assigned to each item. Transactions are created by converting files into transaction sets. Support and confidence are assigned by random experiment to generate precise rules. The occurrence of item is considered for this.

3.2 k-means Clustering

To normalize the continuous data k-means clustering is done. The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the clusters centroid or center of gravity.

K-means clustering is used for grouping. k-means clustering is used to normalize the continuous data. The point of normalization is to make variables comparable to each other. E.g. We can measure temperature in both farhenheit and centigrade. Both are valid but they produce different numbers. Normalization is the process of reducing measurements to a “neutral” or “standard” scale.

3.3 Multilayer Artificial Neural Network

Multilayer Artificial Neural Network is used to increase the accuracy. a neural network is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input topples. Neural network learning is also referred to as connectionist learning due to the connections between units. Neural networks involve long training times.

Three layer neural network is used which consists of input layer, hidden layer and output layer. Back propagation algorithm is used for training of the data. The data (rows) are presented to the network one at a time, and the weights associated with the input values are adjusted each time. Feedforward neural network is used for testing of the data. Sigmoid Activation function is used to introduce nonlinearity in the model. A neural network element computes a linear combination of its input signals, and applies a sigmoid function to the result.

The output of clustering is in between 0, 1, 2, 3, 4, 5. It is given to each input node of Artificial Neural Network. There are 13 input nodes in and 5 output nodes Artificial Neural Network.

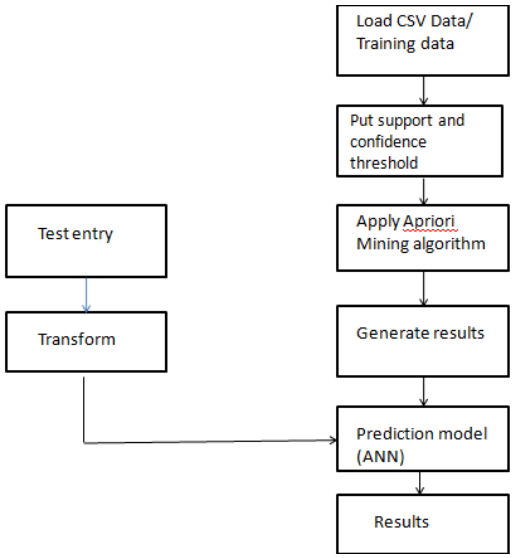

Figure1. System Architecture [1]

3.4 Algorithms

Algorithm 1: Apriori Algorithm

1) Let k=1
2) Generate frequent itemsets of length 1.
3) Repeat until no new frequent itemsets are identified.
    a) Generate length (k+1) candidate itemsets from length k frequent itemsets.
    b) Prune candidate itemsets containing subsets of length k that are infrequent.
    c) Count support of each candidate by scanning DB.
    d) Eliminate candidates that are infrequent.

Algorithm 2: k-means clustering algorithm:

Input

1) K: the number of clusters

2) D: a data set containing n objects

Output: A set of k clusters // Method:

1) Arbitrary choose k objects from D as in initial cluster centers
2) Repeat

3) Reassign each object to the most similar cluster based on the mean value of the objects in the cluster

4) Update the cluster means
5)Until no change

Algorithm 3: Back propagation algorithm

1) First apply the inputs to the network and work out the output.
2) Work out the error for neuron B.

$$Error_B = Output_B(1- Output_B)(T\ target_B - Output_B)$$

3) Change the weight. Let W_{AB}^{+} be the new (trained) weight and W_{AB} be the initial weight.

$$W_{AB}^{+} = W_{AB} + (Error_B * Output_A)$$

4) Calculate the Errors for the hidden layer neurons. Back Propagate them from the output layer.
5) $Error_A = Output_B(1- Output_B)(Error_A\ W_{AB} + Error_C\ W_{AC})$

6) Go to step 3.
Where,
$Error_B$ = Error of neuron B.
W_{AB}^{+} = New weight.
$Error_A$ = Error for hidden layer neuron.

IV. SYSTEM ANALYSIS

Apriori Algorithm is used for rule generation of functional rules. Indices are assigned to each item. Transactions are created by converting files into transaction sets. Support and confidence are assigned by random experiment to generate precise rules. The occurrence of item is considered for this. The k-means algorithm is used for grouping and normalization of continuous data. The point of normalization is to make variables comparable to each other. Firstly grouping of data is done. Then labeling to that data is done. The output of k-means algorithm is given as input to ANN prediction ,Multilayer ANN is used for function approximation and to estimate non-linear relationship among the variables. Three layer neural network is used which consists of input layer, hidden layer and output layer. Backpropagation algorithm is used for training of the data. The data (rows) are presented to the network one at a time, and the weights associated with the input values are adjusted each time. Feedforward neural network is used for testing of the data. Sigmoid Activation function is used to introduce nonlinearity in the model. A neural network element computes a linear combination of its input signals, and applies a sigmoid function to the result.

VI. RESULTS

The predictive accuracy of the derived FARs is compared with the confidence value of the conventional ARs. The output in the dataset is compared with output of ANN. The same entries shows accuracy
and are denoted as positive entries. The rest of the entries are denoted as negative entries. There is tradeoff between accuracy and time. The accuracy of rule generation is increased but the time complexity is also increased. Accuracy is calculated by the following formula:

Accuracy= total true count/ total size of data set

The threshold for support is 40 and threshold for confidence is 70. This threshold gives precise rules.

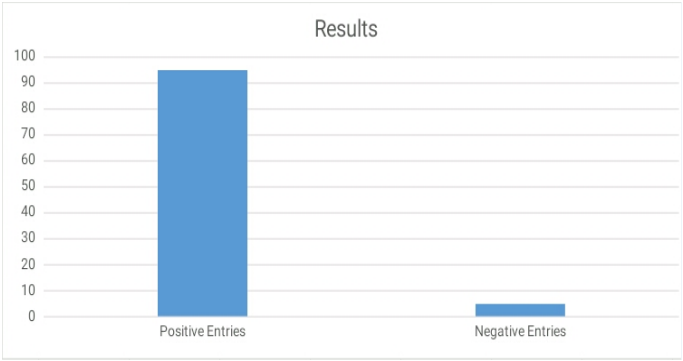

Figure 2: Results

VII. CONCLUSION

An alternative representation of ARs allows for inclusion of nonlinearity, continuous variables without discretization, and multiple RHS. A novel form of association rules (ARs) do not require discretization of continuous variables or the use of intervals in either sides of the rule. This rule form captures nonlinear relationships among variables, and provides an alternative pattern representation for mining required relations hidden in a given data set. The contribution is to convert the ANN into deep ANN using deep learning concept. This will improve the accuracy of FAR mining.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, Mining association rules between sets of items in large databases, in Proc. ACM SIGMOD Int. Conf. Manage. Data, 1993.

[2] .Pahlavan,Feritakgul and Y.YE,Taking Positioning Indoor Wi-Fi Local-ization and GNSS, Inside GNSS Journal May 2010: 40-47.

[3] B. Lent, A. Swami, and J. Widom, Clustering association rules, in Proc. 13th Int. Conf. Data Eng., Apr. 1997.

[4] Z.Dong, Y.Wu and D.Sum,Data Fusion of the Real-Time Positioning System based on RSSI and TOF, 5th International Conference on Intelligent Human-Machine Systems and Cybernetics,2013:503-506.

[5] J. Mata, J. L. Alvarez, and J. C. Riquelme, An evolutionary algorithm to discover numeric association rules, in Proc. ACM Symp. Appl. Comput., 2002.

[6] B. Alatas, E. Akin, and A. Karci, MODENAR: Multi-objective differ-ential evolution algorithm for mining numeric association rules, Appl. Soft Comput., vol. 8, no. 1, pp. 646656, 2008.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, AI Mag., vol. 17, no. 3, pp. 3755, 1996.

[8] R. Agrawal and R. Srikant, Fast algorithms for mining association rules in large databases, in Proc. 20th Int. Conf. Very Large Data Bases, vol. 1215. 1994.

[9] A. Savasere, E. Omiecinski, and S. B. Navathe, An efcient algorithm for mining association rules in large databases, in Proc. 21st Int. Conf. Very Large Data Bases, 1995.

[10] H. Toivonen, Sampling large databases for association rules, in Proc. 22nd Int. Conf. Very Large Data Bases, 1994.

[11] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, Discovering frequent closed itemsets for association rules, in Proc. 7th Int. Conf. Database Theory, 1999.

[12] J. Han, J. Pei, and Y. Yin, Mining frequent patterns without candidate generation, ACM SIGMOD Rec., vol. 29, no. 2, pp. 112, 2000.

[13] J. Han, H. Cheng, D. Xin, and X. Yan, Frequent pattern mining: Current status and future directions, Data Mining Knowl. Discovery, vol. 15, no. 1, pp. 5586, Aug. 2007.

[14] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, Dynamic itemset counting and implication rules for market basket data, ACM SIGMOD Rec., vol. 26, no. 2, pp. 255264, 1997.

[15] A. T. H. Sim, M. Indrawan, S. Zutshi, and B. Srinivasan, Logicbased pattern discovery, IEEE Trans. Knowl. Data Eng., vol. 22, no. 6, pp. 798811, Jun. 2010.

[16] W.-H. Au, K. C. C. Chan, and X. Yao, A novel evolutionary data mining algorithm with applications to churn prediction, IEEE Trans. Evol. Comput., vol. 7, no. 6, pp. 532545, Dec. 2003.

[17] Y.-H. Wu, C.-M. Chiang, and A. L. P. Chen, Hiding sensitive association rules with limited side effects, IEEE Trans. Knowl. Data Eng., vol. 19, no. 1, pp. 2943, Jan. 2007.

[18] D. Brauckhoff, X. Dimitropoulos, A. Wagner, and K. Salamatian, Anomaly extraction in backbone networks using association rules, in Proc. 9th ACM SIGCOMM Conf. Internet Meas. Conf., 2009.

[19] D. K. Y. Chiu and T. W. H. Lui, NHOP: A nested associative pattern for analysis of consensus sequence ensembles, IEEE Trans. Knowl. Data Eng., vol. 25, no. 10, pp. 23142324, Oct. 2013.

[20] T. Jiang, A.-H. Tan, and K. Wang, Mining generalized associations of semantic relations from textual Web content, IEEE Trans. Knowl. Data Eng., vol. 19, no. 2, pp. 164179, Feb. 2007.

[21] A. Salleb-Aouissi, C. Vrain, and C. Nortet, QuantMiner: A genetic algorithm for mining quantitative association rules, in Proc. Int. Joint Conf. Artif. Intell., 2007.

[22] K. Taboada, E. Gonzales, K. Shimada, S. Mabu, K. Hirasawa, and J. Hu, Association rule mining for continuous attributes using genetic network programming, IEEJ Trans. Elect. Electron. Eng., vol. 3, no. 2, pp. 199211, Mar. 2008.

[23] H. Ishibuchi, T. Nakashima, and T. Yamamoto, Fuzzy association rules for handling continuous attributes, in Proc. IEEE Int. Symp. Ind. Electron., vol a, 2001.

[24] S. G. Matthews, M. A. Gongora, and A. A. Hopgood, Evolutionary algorithms and fuzzy sets for discovering temporal rules, Int. J. Appl. Math. Comput. Sci., vol. 23, no. 4, pp. 855868, 2013.

[25] Y. Aumann and Y. Lindell, A statistical theory for quantitative associa-tion rules, J. Intell. Inf. Syst., vol. 20, no. 3, pp. 255283, 2003.

[26] G. I. Webb, Discovering associations with numeric variables, in Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2001, pp. 383388.

[27] U. Rckert, L. Richter, and S. Kramer, Quantitative association rules based on half-spaces: An optimization approach, in Proc. 4th IEEE Int. Conf. Data Mining, Nov. 2004.

[28] E. Georgii, L. Richter, U. Rckert, and S. Kramer, Analyzing microarray data using quantitative association rules, Bioinformatics, vol. 21, no. 2, pp. ii123ii129, 2005.

[29] G. Cybenko, Continuous valued neural networks with two hidden layers are sufcient, Univ. Illinois UrbanaChampaign, Champaign, IL, USA, Tech. Rep., 1988.

[30] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control, Signals, Syst., vol. 2, no. 4, pp. 303314, 1989.

[31] K. Hornik, M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators, Neural Netw., vol. 2, no. 5, pp. 359366, 1989.

