Volume 4, Issue 3, March 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Fuzzy Driven Data Mining Techniques for Effective Prediction

S.Palani Murugan*, A.Javed Sultan, V.M.Suresh, N.Murali
Department of Information Technology
E.G.S.Pillay Engineering College, Nagapattinam
India

A.Ramesh Kumar
Department of Information Technology
Sree Sowdambika College of Engineering, Arupukottai
India

Abstract— In today's computing era, data mining has been the one that turns all computing probable into possible, by means of prediction, description and understanding. For achieving it all, various data mining algorithms are being in use. Fuzzy in simple is a form of many valued logic and it's prominent for its wide applicability. This paper surveys the applications of fuzzy in data mining analysis. Put differently, fuzzy can be used for improving the efficiency otherwise for providing some kind of optimization to the process for achieving better results. Now it's straightforward to reveal that the data mining algorithms that are ambitious about fuzzy computing results in improved data analysis. The below sections introduces data mining and fuzzy and most essentially about the impact of fuzzy in data analysis/extraction.

Keywords— Fuzzy, Classification, Clustering, Association rule mining, Outlier analysis

I. INTRODUCTION

Data mining is promising for its wide range of applications, and changed the course of data extraction techniques over the period of time. Data mining is known for its effectiveness. The amount of data in the world and in our lives seems ever increasing and there is no end in sight. In data mining, data is stored electronically and the search is automated by computer. Even this is not particularly new, Economists, statisticians, forecasters, and communication engineers have long worked with the idea that patterns in data can be sought automatically identified, validated, and used for prediction. Data mining is about solving problems by analysing data already present in databases.

This paper aims at providing a special look at data mining algorithms those have been influenced by fuzzy techniques and how well the effectiveness or accuracy of one particular algorithm is improved given the same testing conditions.

II. BASELINE

The data mining techniques [1] that play vital role in Decision Support Systems (DSS) and which are influenced by fuzzy techniques are taken into consideration, namely Association Rule Mining, Classification, Clustering and Outlier Analysis. Decision support systems enable business and organizational decision-making activities in a more balanced way. With the evolving nature of information and data, it is challenging to extract key information from the large data set, which is most valuable during decision-making. In order to develop systems for digging down the data warehouse, data mining techniques are useful at the very core of such systems. Already many data mining algorithms have been employed enough in DSS and results in tremendous popularity. Still there is need for soft computing techniques like fuzzy logic for data mining algorithms before they are applied in DSS.

III. FUZZY LOGIC IN MACHINE LEARNING

Earlier machine learning systems were very dominant in computing environment. Data mining can be referred as an advancement of online analytical processing (OLTP). So data analysis can effectively be accomplished by using data mining techniques. As data mining is a multi-disciplinary field, it can assisted by statistics, signal processing and information science etc. A normal machine learning system can do data analysis effectively but the problem is that it cannot handle large amount of data. To handle a huge amount of data, a system needs to be scalable and efficient algorithmic techniques. So the gap/problem of handling a large volume of data can be overcome if data mining techniques are used for data analytics. Soft computing is another trait that may be used with data mining techniques to improve its effectiveness. Fuzzy logic, neural networks, rough set theory and evolutionary algorithms like genetic algorithms are coming under the category of soft computing techniques. Even though data mining techniques were good enough to be effective on high dimensional data and large volume of data, it still needed to incorporate some other techniques to make it more reliable and scalable. So this specific need was convinced by using soft computing techniques on data mining algorithms. Techniques like fuzzy logic and neural networks can be applied to data analysis process along data mining algorithms. Fuzzy logic plays vital role in improving the effectiveness and to increase the efficiency of various data mining techniques like association rule mining, clustering, classification, sequential pattern mining, regression analysis and outlier analysis. We insist on the role and impact of fuzzy systems [2] on classification, clustering, association rule mining and finally outlier analysis in specific.

IV. IMPACT OF FUZZY SYSTEMS IN DATA MINING ALGORITHMS

The below sections describe the role of fuzzy systems in data mining algorithms and more importantly how closely it is used for optimizing the results produced by lone data mining techniques.

A. Classification

Classification is the process of deriving a model or function from the given set of data which includes the result and the derived model can be used then for predicting the result of a data which has no result associated with it. It is normally a two steps process namely training and testing. The former is known for deriving a model the given dataset and the latter is known for applying or using the model for prediction. The model can be of a set of if...then rules that forms a classifier in the end. Classifier is like container which has a set of rules (if...then). Typical application of classification includes finding risk customer, predicting heart disease etc. The main objective of introducing fuzzy in classification process [3] [4] [10] is to convert the objectively measurable parameters to category memberships. The membership ranges from 0.0 to 1.0. A number of tools which are capable of generating this membership values are in use still it is upto us to define the linguistic terms. For an instance, a person's age can be an attribute/parameter in a data analysis task, age is a general attribute that can be categorised as young, middle and old. Each and every term categorised has a range of values specified. Consider the following set of values for the terms young [15-30], middle [31-55] and old [56-75]. Each term categorised here can be referred to as linguistic terms.

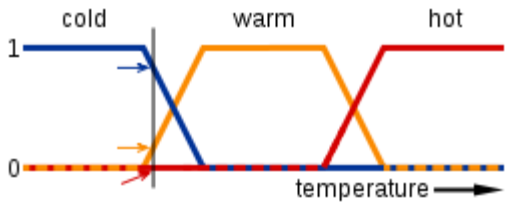

Fig. 1. Membership values for attribute "temperature"

The above diagram shows the membership chart for an attribute "temperature", categorised in to three terminologies, referred to as linguistic terms, cold, warm and hot that typically range from 0 to 1. Classification process uses the membership function that tells whether a value is a member of a class. Generally 0-indicates no membership and 1-indicates full membership.

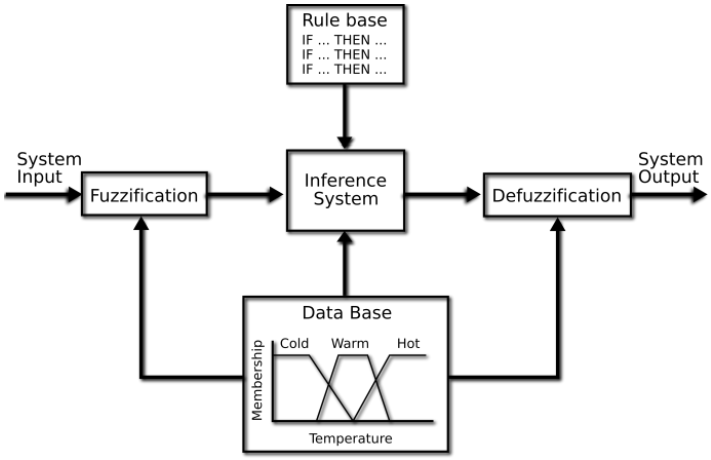

Fig. 2. Data flow in a fuzzy system

The advantage of using fuzzy in classification is interpretability, mainly when it comes to high dimensional data, the antecedent part might not have all the attributes in it. More importantly fuzzy normalizes the attributes before training process. So it fancies the chances of getting compact and accurate rules for classification process.

B. Association Rule Mining

Association rule mining is the process of finding the co-occurrences within a set of data. Sometimes it is referred to as frequent pattern mining. Frequent itemset is a one that appears frequently together in a transaction. Finding such frequent data has been useful in finding the relationship among data. Typical application of association rule mining includes market basket analysis in which knowing customers' buying behaviour. It is helpful for retailers to change the marketing strategies according to customers' buying behaviour to improve their marketing and to develop business. Support and confidence are the two measures used for finding the interestingness of the rule. It can be used as threshold values, which are normally set by domain experts.

In the earlier part of data mining techniques especially in association rule mining only binary valued logic had been used. It's very easier to form as bitmap approach would be handy enough to find the equivalent binary values. Nowadays

transaction data are very much in form of qualitative and fuzzy. So there is a need to design techniques that are capable of handling various forms of data. It leads to the desperate need to bring fuzzy systems in to play. As described in section IV A, fuzzy is very predominant nowadays, as to be able to improve data mining systems considerably. The quantitative approach allows an item either to be member of an interval or not. This leads to an under or over estimation of values that are close to the borders of such crisp sets. To get rid of this problem, the approach of fuzzy association rules [5] can be used. It leaves the intervals to overlap, making the set fuzzy instead of being crisp. So it can be of partial membership to more than one set. This partial membership effectively overcomes the problem of "sharp boundary problem".

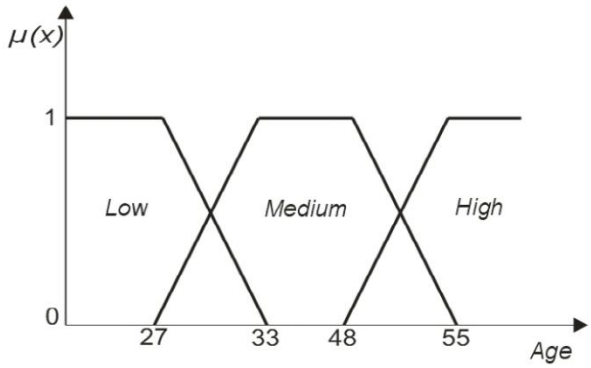

Fig. 3. Fuzzy partition of attribute "age"

In contrast to the one defined in section IV.A, we define young [0-33], middle [27-55] and old [48-75] leads to the partial membership concept rather than to have full or no membership. Considering age value 30 might fall in young as well as middle despites the sum will always be 1.0 referring partial membership. It adds a greater flexibility to define more accurate rules.

C. Clustering

Clustering is the process of finding groups from the given set of data that highly depends on similarity. In contrast to classification (supervised learning), it is un-supervised learning. The term un-supervised learning means that the result or class label for the records in the given dataset may not be known. Considering this case, clustering will be hugely effective when the class label is unknown. The outcome of clustering as a process is always a set of groups/clusters where intra-cluster values are highly similar and inter-cluster values are less similar. Hard clustering requires that an object either does or does not belong to a cluster or a group. The problem is that an object may be a part of more than one group. The objective of using fuzzy in clustering will get rid of the limitations of partitioning data into a specified number of mutually exclusive sets (hard clustering). Fuzzy clustering [6] [7] [8] methods allow the objects to belong to several groups simultaneously. The fuzzy membership value plays a vital role in defining an objects' part in a cluster(s). An objects' belonging in different clusters specified with different degrees of membership.

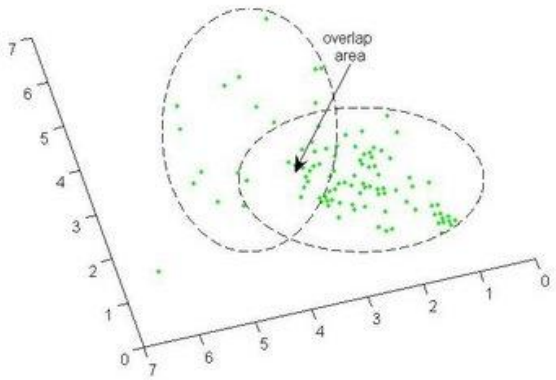

Fig. 4. Object being part of two clusters

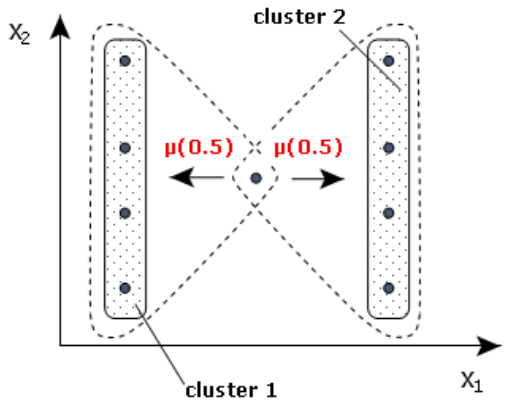

Fig. 5. Fuzzy clusters with respective membership

The above figures are representatives for the concept of an object can be a member of many clusters and the membership values for the object. Fuzzy systems used for clustering improves the accuracy as well as the flexibility of using objects.

D. Outlier Analysis

A dataset may contain data object that may not fit with the general behaviour of the data. Such data objects are referred as outliers. In most data analysis these outliers are left apart. But it needs to be considered for data analysis. Such process is referred to as outlier mining. Outlier mining is a two steps process namely detecting the outlier and

handling/treating outliers. Outlier detection is a tough one in nature. Earlier a record which has an outlier will be completely discarded. Outliers can be detected by some statistical methods like linear regression model. The detection scheme may be supervised, unsupervised or semi-supervised. Outlier detection [9] is applicable in various domains such as fraud detection, intrusion detection etc.

To decide an outlier it needs a lot of human work. However the human perception of an outlying value depends on some additional factors, like the number of records, where the value occur, total number of non null value etc. To automate the process of detecting outliers, we represent the conformity of an attribute value by a well defined membership function. The conformity [11] becomes very close to zero, when the number of value occurrences is much smaller than the average number of records per value. If the data set is small, a single occurrence of a value is considered an outlier. This is how fuzzy systems are very effective in finding outliers.

V. CONCLUSIONS

The advantages of using fuzzy in various data mining techniques have been given here. We considered only the predominant techniques like clustering, classification, association and outlier mining. But combination of these techniques results in associative classification, association in clustering, classification includes clustering (semi supervised) has been already proposed. More importantly, fuzzy membership is spoken throughout the paper. The counterpart is also there in name of weighted fuzzy systems. It is up to the domain expert to go for which type of fuzzy system based on the requirement and the type of data being used. But the main goal behind writing this paper is to give the core idea and facts of improvements in data mining algorithms and various machine learning techniques especially when these are adapted to be influenced by fuzzy systems. There has been a lot of research going on in the area of soft computing in data mining and so many computational methods/models like neuro-fuzzy, genetic-fuzzy and rough set theories have been proposed in the recent past. But this paper is aimed at providing the impacts and role of fuzzy in optimizing or improving data mining techniques.

ACKNOWLEDGMENT

The paper would not have been possible without the efforts from some incredible authors on data mining systems especially in clustering, classification, association rule mining and outlier mining. I have learnt about fuzzy systems with very minor impact as it was very vast in nature. My special mention for the authors and experts who are nurtured with the knowledge of fuzzy systems especially from whom i have learnt about fuzzy. Last but not least, final mention for my parents, friends, colleagues and my teachers of all time.

REFERENCES

[1] Jiawei Han, Micheline Kamber, *Data Mining Concepts and Techniques*, 2nd ed., The Morgan Kaufmann Series in Data Management Systems

[2] Rudolf Kruse, Detlef Nauck and Christian Borgelt., *Data Mining with Fuzzy Methods: Status and Perspectives*, Proceedings of the EUFIT'99, 1999.

[3] F.Okeke and A.Karnieli, "*Methods for fuzzy classification and accuracy assessment of historical aerial photographs for vegetarian change analyses.Part.I:Algorithm development* ," International journal of remote sensing., vol. 27, pp. 153-176, Jan. 2006.

[4] Andreas Meier, Nicolas Werro, Martin Albrecht and Miltiadis Sarakinos, "*Using a Fuzzy Classification Query Language for Customer Relationship M anagement*," in *Proc.of the 31st VLDB Conference*, pp. 1089-1096.

[5] Bakk. Lukas Helm, "*Fuzzy Association Rules, An implementation in R*," Master Thesis, Vienna University of Economics and Business Administration, Vienna, 2-8-2007.

[6] Sadina Gagula-Palalic, "*Fuzzy Clustering models and algorithms for pattern recognition*", Sarajevo 2008.

[7] Babuska, "Fuzzy Clustering", pp 55-72

[8] M.S Yang, "*A Survey of fuzzy clustering*", Mathl. Comput. Modelling Vol. 18, No. 11, pp 1-16, 1993.

[9] Nilam Upasani, Hari Om, "*Outlier Detection: A Survey on Techniques Involving Fuzzy and Neural Approaches*", IEEE Workshop on Computational Intelligence, July 2013, pp 28-32

[10] Johannes A.Roubos, Magne Setnes and Janos Abonyi, "*Learning fuzzy classification rules from labeled data*", Information sciences 150 (2003) 77-93.

[11] Mark Last and Abraham Kandel, "*Automated Detection of Outliers in Real-World Data*", pp 1-10

