ISSN: 2321-7782 (Online)

Volume 2, Issue 12, December 2014

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Role of Fuzzy Logic in Data Mining

M.Rekha[1]
Computer Science and Engineering
Siddharth Institute of Engg & Tech
Puttur, India.

M.Swapna[2]
Computer Science and Engineering
Shree Institute of Technical Education
Tirupati, India.

Abstract: This paper focuses on real-world applications of fuzzy techniques for data mining. It gives a presentation of the theoretical background common to all applications, lying on two main elements: the concept of similarity and the fuzzy machine learning framework. It then describes a panel of real-world applications covering several domains namely medical, educational, chemical and multimedia. There are two main types of uncertainty in supervised learning: statistical and cognitive. Statistical uncertainty deals with the random behavior of nature and all existing data mining techniques can handle the uncertainty that arises (or is assumed to arise) in the natural world from statistical variations or randomness. Cognitive uncertainty, on the other hand, deals with human cognition.

Keywords: data mining, fuzzy logic, fuzzy set theory, Inference, linguistic variables.

I. Introduction

Data mining: Data mining is also named as "Knowledge mining" from data. Data mining is an essential process where intelligent methods are applied to extract data patterns [1]. Data mining is a process that analyzes large amounts of data to find new and hidden information. In other words; it is the process of analyzing data from different perspectives and summarizing it into some useful information.

The following are the different data mining techniques:

- Association
- Classification
- Clustering
- Sequential patterns

Fuzzy logic:

We begin by presenting some of the basic concepts of fuzzy logic. The main focus, however, is on those concepts used in the induction process when dealing with data mining.

Fuzzy Logic is a logic system for reasoning that is approximate rather than exact.The fundamental unit of a fuzzy logic is the fuzzy set. Given the universal set X in order to define a fuzzy set A on X, we define a membership function A: X -> [0,1] that maps element x of X into real numbers in [0,1].A(x) is interpreted as the degree to which x belongs to the fuzzy set A. We sometimes write fuzzy set A as $\{(x, A(x))| \ x \in X\}$.

In classical set theory, a certain element either belongs or does not belong to a set. Fuzzy set theory, on the other hand, permits the gradual assessment of the membership of elements in relation to a set.

Let U be a universe of discourse, representing a collection of objects denoted generically by u. A fuzzy set A in a universe of discourse U is characterized by a membership function μ_A which takes values in the interval [0, 1]. Where $\mu_A(u) = 0$ means that u is definitely not a member of A and $\mu_A(u) = 1$ means that u is definitely a member of A.

The main difference between classical set theory and fuzzy set theory is that the latter admits to partial set membership. A classical or crisp set, then, is a fuzzy set that restricts its membership values to {0, 1}, the endpoints of the unit interval.

Linguistic variables: A linguistic variable such as *age* may have a value such as *young* or its antonym *old*. However, the great utility of linguistic variables is that they can be modified via linguistic hedges applied to primary terms. The linguistic hedges can be associated with certain functions.

Applications Of Fuzzy Logic In Data Mining

This chapter focuses on real-world applications of fuzzy techniques for data mining. It first gives a brief presentation of the theoretical background common to all applications (Sect. 2), decomposed into two main elements: the notion of similarity and the fuzzy machine learning techniques that are applied in the described applications. Indeed, similarity, or more generally comparison measures are used at all levels of the data mining and information retrieval tasks: at the lowest level, they are used for the matching between a query to a database and the elements it contains, for the extraction of relevant data. Then similarity and dissimilarity measures can be used in the process of cleaning and management of missing data to create a training set. In the various techniques to generalize particular information contained in this training set, dissimilarity measures are used in the case of inductive learning, similarity measures for case-based reasoning or clustering tasks. Eventually, similarities are used to interpret results of the learning process into an expressible form of knowledge, for instance through the definition of prototypes. Section 2.1 presents the similarity notion more formally. Section 2.2 considers a complementary component of similarity, the fuzzy learning techniques in which they can be used. It describes methods used in the applications presented in Sect. 3, namely fuzzy decision trees that perform fuzzy inductive learning, fuzzy prototype extraction that provides flexible.

Real-World Fuzzy Logic Applications

Characterization of data sets and fuzzy clustering that identifies relevant subgroups in data sets. Finally we describe real-world applications exploiting these methods and belonging both to the data mining and information retrieval fields. They cover several domains, such as medical, educational, chemical and multimedia.

II. Theoretical Background

In this section, we recall the theoretical background common to the applications presented in Sect. 3, considering successively the notion of similarity (Sect. 2.1) and fuzzy machine learning techniques (Sect. 2.2).

2.1 Similarity

The notion of similarity or more generally of comparison measures, is central for all real-world applications: it aims at quantifying the extent to which two objects are similar, or dissimilar, one to another, providing a numerical value for this comparison. Similarities and dissimilarities between objects are generally evaluated from values of their attributes or variables characterizing these objects. It is the case in various domains, such as statistics and data analysis, psychology and pattern recognition for instance. Dissimilarities are classically defined from distances. Similarities and dissimilarities are often expressed from each other: the more similar two objects are, the less dissimilar they are, the smaller their distance. Weights can be associated with variables, according to the semantics of the application or the importance of the variables. It appears that some quantities are used in various environments, with different forms, based on the same principles. For instance, the most classic dissimilarity measures between two objects with continuous numerical attributes are the Euclidian distance, the Manhattan distance, and more generally Minkowski distances. In the case of binary attributes, coefficients introduced by Russel-Rao, Jaccard, Dice or Ochiai are very popular.

2.2 Fuzzy Machine Learning

The second part of the theoretical background common to all applications described in Sect. 3 concerns the fuzzy machine learning techniques that use the previous similarity measures. Machine learning is an important way to extract knowledge from sets of cases, especially in large scale databases. In this section, we consider only the fuzzy machine learning methods that are used in the applications described in Sect. 3, leaving aside other techniques as for instance fuzzy case based reasoning or fuzzy association rules (for a complete review on fuzzy learning methods, the interested reader is referred to [2]).Three methods are successively considered: fuzzy decision trees, fuzzy prototypes and fuzzy clustering. The first two belong to the supervised learning framework, i.e. they consider that each data point is associated with a category. Fuzzy clustering belongs to the unsupervised learning framework, i.e. no a priori decomposition of the data set into categories is available.\

a. Fuzzy Decision Trees

Fuzzy decision trees (FDT) are particularly interesting for data mining and information retrieval because they enable the user to take into account imprecise descriptions of the cases, or heterogeneous values (symbolic, numerical, or fuzzy) [3, 4, 5, and 6]. Moreover, they are appreciated for their interpretability, because they provide a linguistic description of the relations between descriptions of the cases and decision to make or class to assign [3]. The rules obtained through FDT make it easier for the user to interact with the system or the expert to understand, confirm or amend his own knowledge. Another quality of FDT is their robustness, since a small variation of descriptions does not drastically change the decision or the class associated with a case, which guarantees a resistance to measurement errors and avoids sharp differences for close values of the descriptions.

b. Fuzzy Prototype Construction

Fuzzy prototypes [7, 8, 9] constitute another approach to the characterization of data categories: they provide descriptions or interpretable summarizations of data sets, so as to help a user to better apprehend their contents: a prototype is an element chosen to represent a group of data, to summarize it and underline its most characteristic features. It can be defined from a statistic point of view, for instance as the data mean or the median; more complex representatives can also be used, as the Most Typical Value [10] for instance. The prototype notion was also studied from a cognitive science point of view, and specific properties were pointed out [11]: it was shown that a prototype underlines the common features of the category members, but also their distinctive features as opposed to other categories, underlining the specificity of the group. Furthermore, prototypes were related to the typicality notion, i.e. the fact that all data do not have the same status as regards the group: some members of the group are better examples, more representative or more characteristic than others. It was also shown [11] that the typicality of a point depends both on its resemblance to other members of the group (internal resemblance), and on its dissimilarity to members of other groups (external dissimilarity). These definitions were exploited by Rifqi, who proposed a construction method implementing these principles and exploiting the similarity measure framework presented in Sect. 2.1. More precisely, the method consists in first computing internal resemblance and external dissimilarity for each data point they are respectively defined as the aggregation (mean or median e.g.) of the resemblance to the other members of the group, and as the aggregation of the dissimilarity to members of other groups, for a given choice of the resemblance and dissimilarity measures (see Sect 2.1). In a following step, a typicality degree is computed for each data point as the aggregation of its internal resemblance and external dissimilarity. In a last step, the prototype itself is defined, as the aggregation of the most typical category members.

III. Real World Applications

In this section, a panel of real-world fuzzy logic applications is presented, based on the similarity framework and the fuzzy learning methods described in the previous section. They belong both to the data mining and information retrieval fields, and cover several domains, namely medical, educational, and multimedia. For each application, the objective of the task, the considered data, the applied method and the obtained results are successively described.

3.1 Medical Applications

Medical applications are good cases where Fuzzy Set Theory can bring out enhancement as compared to classic algorithms because most of the attributes used here to characterize cases are associated with imprecise values. In this section, we present three applications of data mining, respectively to prevent cardio-vascular diseases, to measure asthma severity and to detect malign micro calcifications in mammographies.

Data Mining to Prevent Cardio-Vascular Diseases: This project was done thanks to financial supports by INSERM and was led by M.-C. Jaulent (INSERM ERM 0202). Researchers from several French universities collaborated with a medical scientist on a well-known database to prevent cardio-vascular diseases.

Objective: The main objective here was to find discriminating features in order to prevent cardio-vascular diseases. Predictions should help medical scientists to detect and prevent cardio-vascular diseases for hypertensive patients.

3.2 Educational Applications

Providing Interpretable Characterizations of Students: In this section, we consider another domain application for fuzzy machine learning methods, namely the educational domain. The presented application was performed in the framework of a project with the schoolbook publisher Bordas-Nathan.

Objective: The considered task consists in characterizing students, through the identification of relevant groups of students having the same characteristics, and the comparison of several student classes, to determine whether the classes present the same characteristics or not. Of special importance is the interpretability of the results, to enable a teacher to exploit the information and the structure identified in the student data.

3.3 Multimedia Applications

In this section, we consider applications in data mining and information retrieval in the multimedia field. We describe first an image retrieval application based on a visual similarity navigation paradigm and second a learning approach for a semantic annotation of a video signal, based on some examples.

Searching in a Clothes Catalogue by Visual Similarity: This work is part of the results of the ITEA European project: KLIMT -KnowLedge InterMedation Technology. Although in this project several laboratories and industry partners were involved to accomplish what follows, the company Sinequa played a significant role.

Objective: The main objective of this work was to enhance a classic text search engine of an on-line clothes catalogue, with an image search tool. A prototype was developed that illustrated the complementarities between the two navigation schemas: text queries and visual-similarity browsing.

IV. CONCLUSION

In this paper we briefly present first two of these essential pillars: fuzzy comparison measures and fuzzy machine learning. Then, based on these two strongly interrelated bases, a set of applications in domains ranging from medical to educational but also chemical and multimedia domains illustrate specific solutions. All these applications focus on information retrieval or data mining, which are in fact, as we saw in this chapter, two components of a same challenge: the search or the extraction of information and knowledge from large amounts of data. For each of the presented solutions, we focused only on how the theory supported the application, ignoring a large set of other difficulties, appearing when dealing with real world challenges: as for instance technical issues, solutions for fast execution (essential in the case of large data sets), management of large data bases, corrupted data, etc. For more details on each of these applications please refer to the corresponding publications. Finally, what makes all these applications unique is the use of fuzzy logic.

References

1. Jiawei Han,Micheline Kamber,Jian Pei, "Data mining concepts and techniques",2012
2. E.Hullermeier Fuzzy methods in data mining and data mining: status and prospects. Fuzzy Sets and Systems, 156(3):387-406, 2005.
3. C.Z Janikow .Fuzzy decision trees:issues and methods IEEE transactions on systems Man and Cybernetics, 28(1):1-14, 1998.
4. M.Ramdani Une approche floue pour traiter les valeurs numeriques en apprentissage, In Journées Francophones d'apprentissage et d'explication des connaissances, 1992.
5. R.Weber Fuzzy ID3: A class of methods for automatic knowledge acquisition. In IIZUKA'92 Proceedings of the 2nd International Conference on Fuzzy Logic, pages 265-268, 1992.
6. Y.Yuan and M.J Shaw.Introduction of fuzzy decision Trees.Fuzzy sets and systems, 69:125-139, 1995.
7. M.-J Lesot,L.Mouillet and B.Bouchon-Meunier Descriptive Concept extraction with exceptions by hybrid clustering
8. M.Rifqi Constructing prototypes from large databases,In Proc. of IPMU'96,1996
9. L.A Zadeh.A note on prototype theory and fuzzy sets.cognition, 12:291-297, 1982.
10. M.Friedman and M.Ming and A.Kandel. On the theory of typicality,International journal of Uncertainity, Fuzzyness and Knowledge-Based Systems, 3(2):127-142, 1995.
11. E.Rosch and C.Mervis .Family Resemblance: Studies of internal structure of categories, Cognitive psychology, 7:573-605, 1975.

AUTHOR(S) PROFILE

M. Rekha received B.Tech degree in Computer Science and Engineering from Sri Venkatesa Perumal College of Engineering and Technology, JNTUA University, Anantapuram, A.P, India in 2009 and completed M.Tech, Computer Science and Engineering, from Sri Venkateswara University College of Engineering, TIRUPATI, A.P, India.Working as assistant professor in the department of CSE in Siddharth Institute of Engineering & Technology, Puttur. Interested areas are Data Mining, Artificial Neural Networks. Attended two National Conferences during 2012 and 2014 and published an International Journal.

M. Swapna received B.Tech degree in Computer Science and Engineering from Gokula Krishna College of Engineering, JNTUA, Anantapur, A.P, and India in 2011 and completed M.Tech in Computer Science and Engineering from Sri Venkateswara University College of Engineering, Tirupati, A.P, and India.Working as assistant professor in the department of CSE in Shree Institute of Technical Education, Tirupati. Interested areas are Data Mining and Fuzzy Logic. Attended Two National Conferences during 2012 and 2014. And published an International Journal.