International Journal on Applications in Information and Communication Engineering
Volume 1: Issue 11: November 2015, pp 8-10. www.aetsjournal.com ISSN (Online) : 2394-6237

A Survey on Incomplete Pattern Classification Method

Amrutha Sree V.M, J. Prakash

***Abstract*—** In this paper, we discuss about various incomplete pattern classification and evidential reasoning techniques used in the area of data mining. 10 papers are taken for survey. Its advantages and disadvantages are discussed.

***Keywords*—** Belief function, credal classification, evidential reasoning, incomplete pattern, missing data

I. INTRODUCTION

Data mining is the computational process of discovering patterns in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use

The categorization of incomplete patterns along with the missing values is a prominent issue in the area of machine learning approach. The missing information data in an incomplete pattern have diverse assessments and the classification of pattern outcomes along with various evaluations might be different. Those uncertainties of categorization are primarily caused through the loss of data in the missing information. To avoid this issue, prototype-based credal classification (PCC) method is used, which can deal with incomplete patterns. The belief function structure utilized typically in evidential reasoning method. The class prototypes acquired through training examples are correspondingly utilized to compute the absent values. Since all these separate categorization results are likely admissible, to unite them all together to achieve the final classification of the incomplete pattern.

A. Pattern classification with missing data: a review

In this paper pattern classification methods are utilized for the applications such as biometric identification, text categorization or medical analysis. Missing or unknown data are a universal problem that prototype detection methods need to handle with when resolving real-time classification tasks. Machine learning schemes [1] and techniques introduced from arithmetic learning premise have been mainly considered and utilized in this area under discussion. Missing data imputation and model based procedure is used for handling missing data. The objective of this research is to examine the missing data issue in model categorization tasks, and to recap as well as evaluate some of the standard techniques utilized for dealing the missing values. However it has issue with provision of inappropriate results for some other applications.

B. Imputing missing values: the effect on the accuracy of classification

This paper proposes linear regression [2] methods that are suggested for efficient and accurate classification. Once the model is constructed, synthetically created missing values would be substituted with imputed values by using mean substitution and regression imputation methods. The result on the precision of the calculations by using models with assigned values has been established through evaluation of the re-classifications using imputed data with the actual incidence or non-occurrence of a succeeding morbid occurrence. This method is used to predict better categorical or numerical values

C. Supervised learning from incomplete data via an EM approach

In this paper supervised learning method introduces the Expectation Maximization (EM) [3] approaches to handle the missing and incomplete datasets. In this research, the framework depends on maximum likelihood density computation for learning from those datasets. EM is utilized for both the estimation of mixture objects and for coping with missing information data. This kind of result algorithm is suitable for extensive array of supervised and unsupervised machine learning problems. However it has issue with speed of the process.

D. Missing data imputation for fuzzy rule-based classification systems

In this paper Fuzzy rule-based classification systems (FRBCSs) are known due to their ability to treat with low quality data and obtain good results in these scenarios. One of the most common methods to overcome the drawbacks produced by missing values is based on pre-processing, formerly known as imputation [4]. From the obtained results, the convenience of using imputation methods for FRBCSs with missing values is stated. The analysis suggests that each type behaves differently while the use of determined missing values imputation methods could improve the accuracy obtained for these methods. Thus, the use of particular imputation methods conditioned to the type of FRBCSs is

Amrutha Sree V.M , PG Scholar, Department of Computer Science and Engineering, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India. (Email: amruthavinod2010@gmail.com)

J. Prakash, Assistant Professor, Department of Computer Science and Engineering, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India. (Email: Jeevaprakash86@gmail.com)

required. The drawback of this paper is occurrence of misclassification results. But it has high search ability to discover quality of fuzzy rules and produces more accurate prediction results.

E. Towards missing data imputation: a study of fuzzy k-means clustering method

This paper present a missing data imputation method based on the most popular techniques in Knowledge Discovery in Databases (KDD), i.e. clustering technique [5]. It combine the clustering method with soft computing, which tends to be more tolerant of imprecision and uncertainty, and apply a fuzzy clustering algorithm to deal with incomplete data. This experiments show that the fuzzy imputation algorithm presents better performance than the basic clustering algorithm. Using this technique efficiency and accuracy is increased and the classifications of results are improved.

Methods for handling missing data can be divided into three categories. The first is ignoring and discarding data, and list wise deletion and pair wise deletion are two widely used methods in this category. The second group is parameter estimation, which uses variants of the Expectation-Maximization algorithms to estimate parameters in the presence of missing data. The third category is imputation, which denotes the process of filling in the missing values in a data set by some plausible values based on information available in the data set.

F. The Combination of Evidence in the Transferable Belief Model

This paper propose evidence concept with belief model which is used to handle the uncertainty issues. It is also producing the recommendations for imprecision values and error values. This scenario built the transferable model without introducing explicitly and implicitly any concept of probability. Dempster's rule [6] of conditioning is one of the natural ingredients of the transferable belief model. In this research, the transferable belief model presents two characteristics: the masses allocation that leads to super additive belief functions to describe someone's degree of belief and a rule to combine two distinct evidences. The interest of the first aspect is usually recognized. But the combination rule was felt to be ad hoc by critics, especially when they interpret the transferable belief model as an upper and lower probabilities model. It is efficiently handling the uncertainty issues and it takes minimum amount of time for execution.

G.Measuring Ambiguity in the Evidence Theory

In this paper, an alternative measure to AU for quantifying ambiguity of belief functions is proposed. This measure, called Ambiguity Measure (AM) [7], besides satisfying all the requirements for general measures also overcomes some of the shortcomings of the AU measure. Indeed, AM overcomes the limitations of AU by: 1) minimizing complexity for minimum number of focal points; 2) allowing for sensitivity changes in evidence; and 3) better distinguishing discord and non specificity. The evidence theory also known as Dempster–Shafer theory is one of the most popular frameworks for dealing with uncertain information. In an equivalent way that Shannon entropy has been used in the probabilities framework, information or preferably uncertainty-based information can be quantified. The issue described in this paper is high computing time complexity. The benefit of this provides consistent results and it increases the efficiency of the system.

H.Classification Using Belief Functions: Relationship between Case-Based and Model-Based Approaches

This paper propose transferable belief model (TBM) to represent quantified uncertainties based on belief functions regardless of any underlying probability model. This shows that both methods actually proceed from the same underlying principle, i.e., the general Bayesian theorem (GBT), and that they essentially differ by the nature of the assumed available information. Prototype based credal classification [8] method is used for incomplete pattern classification method. Here In statistical pattern recognition, two main families of classifiers can be distinguished, namely: 1) methods that directly estimate posterior class probabilities (such as the *k*-nearest neighbor (*k*-NN) rule, decision trees, or multilayer perception classifiers), and 2) methods based on density estimation, in which posterior probability estimates are computed from class conditional densities and prior probabilities using Bayes' theorem. This paper also shows that both methods collapse to a kernel rule in the case of precise and categorical learning data and for certain initial assumptions, and a simple relationship between basic belief assignments produced by the two methods is exhibited in a special case. These results shed new light on the issues of classification and supervised learning in the TBM. It provides less error rate and improves the consistent results.

I. A Neural Network Classifier Based on Dempster-Shafer Theory

In this paper, an adaptive version of this evidence-theoretic classification rule is proposed. In this approach, the assignment of a pattern to a class is made by computing distances to a limited number of prototypes, resulting in faster classification and lower storage requirements. Based on these distances and on the degree of membership of prototypes to each class, basic belief assignments BBA's are computed and combined using Dempster's rule. This rule can be implemented in a multilayer neural network [9] with specific architecture consisting of one input layer, two hidden layers and one output layer. The weight vector, the receptive field and the class membership of each prototype are determined by minimizing the mean squared differences between the classifier outputs and target values. It is used to produce high level classification results and able to deal with uncertainty problems

J. A k-nearest neighbor classification rule based on Dempster-Shafer Theory

In this paper, the problem of classifying an unseen pattern on the basis of its nearest neighbors in a recorded data set is addressed from the point of view of Dempster- Shafer theory [10]. Each neighbor of a sample to be classified is considered as an item of evidence that supports certain hypotheses regarding the class membership of that pattern. The degree of support is defined as a function of the distance between the two vectors. The evidence of the *k* nearest neighbors is then pooled by means of Dempster's rule of combination. This approach provides a global treatment of such issues as ambiguity and distance rejection, and imperfect knowledge regarding the class membership of training patterns. The effectiveness of this classification scheme as compared to the voting and distance-weighted *k*-NN procedures is demonstrated using several sets of simulated and real-world data.

II. CONCLUSION

Missing or incomplete data is a usual drawback in many real-world applications of pattern classification. In this paper, we discussed about various incomplete pattern classification techniques and evidence theory concepts in data mining. But some classification techniques are too costly to implement in real time. The results of these techniques are analyzed. Compared to all these result prototype based credal classification method and belief function gives the better result and is cost effective.

REFERENCES

[1] P. Garcia-Laencina, J. Sancho-Gomez, and A. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Comput. Appl.* vol. 19, no. 2, pp. 263–282, 2010.

[2] D. J. Mundfrom and A. Whitcomb, "Imputing missing values: The effect on the accuracy of classification," *MLRV*, vol. 25, no. 1, pp. 13–19, 1998.

[3] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an EM approach," in *Advances in Neural Information Processing Systems*, vol. 6, J. D. Cowan *et al*., Eds. San Mateo, CA, USA: Morgan Kaufmann, 1994, pp. 120–127.

[4] J. Luengo, J. A. Saez, and F. Herrera, "Missing data imputation for fuzzy rule-based classification systems," *Soft Comput.*, vol. 16, no. 5, pp. 863–881, 2012.

[5] D. Li, J. Deogun, W. Spaulding, and B. Shuart, "Towards missing data imputation: A study of fuzzy k-means clustering method," in *Proc. 4th Int. Conf. Rough Sets Current Trends Comput. (RSCTC04)*, Uppsala, Sweden, Jun. 2004, pp. 573–579.

[6] P. Smets, "The combination of evidence in the transferable belief model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 5, pp. 447 458, May 1990.

[7] A.-L. Jousselme, C. Liu, D. Grenier, and E. Bossé, "Measuring ambiguity in the evidence theory," *IEEE Trans. Syst., Man, Cybern. A, Syst.*

[8] T. Denoeux and P. Smets, "Classification using belief functions: Relationship between case-based and model-based approaches," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 6, pp. 1395–1406, Dec. 2006.

[9] T. Denoeux, "A neural network classifier based on Dempster–Shafer theory," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 30, no. 2, pp. 131–150, Mar. 2000.

[10] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster–Shafer Theory," *IEEE Trans. Syst., Man, Cybern.*, vol. 25, no. 5, pp. 804–813, May 1995.