International Journal on Applications in Information and Communication Engineering
Volume 1: Issue 11: November 2015, pp 24-28. www.aetsjournal.com ISSN (Online) : 2394-6237

A Study On Systematic and Constructive Approaches for Insufficient Pattern Recognition with Evidence Theory Concept

Nisha Nandhini A, Mr. Karthik, Mr.J.Prakash, Dr.P.Ezhilarasu

Abstract— Data mining is also known as Knowledge Discovery in Data (KDD). Data mining techniques are the result of a long process of research. In this paper, we discuss about various incomplete pattern grouping and evidential reasoning techniques used in the area of data mining. Grouping consists of predicting a certain outcome based on a given input. This survey shows the advantages and disadvantages of various technology used in classification.

Keywords— Incomplete data, credal classification, pattern matching, incomplete pattern, KNN.

I. INTRODUCTION

Data Mining is defined as the procedure of extracting information from huge sets of data. Data mining is the computational process of discovering patterns in large data sets ("big data") involving methods at the intersection, machine learning, statistics, and database systems. Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. 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

Classification of data is done with the collection of huge data. The data set is classified using any data classification algorithm. The rule for various classification algorithms must be defined, as shown in figure 1.

The grouping 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

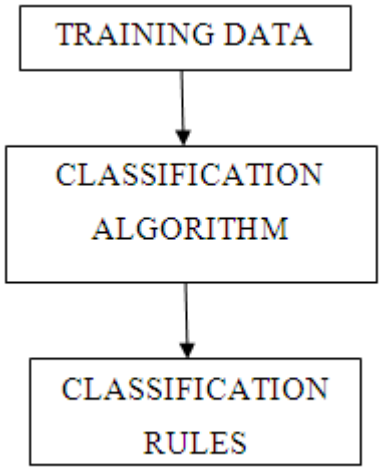

Fig.1.Classification Algorithm

II.. NEAREST NEIGHBOUR ALGORITHM

The nearest neighbor algorithm was one of the first algorithms used to determine a solution for travelling salesman problem.

These are the steps involved in the algorithm:

1. Start on an arbitrary vertex as current vertex.
2. Find out the shortest edge connecting current vertex and an unvisited vertex V.
3. Set current vertex to V.
4. Mark V as visited.
5. If all the vertices in domain are visited, then terminate.
6. Go to step 2.

The nearest neighbour algorithm is easy to implement and executes quickly, but it can sometimes miss shorter routes which are easily noticed with human insight, due to its "greedy" nature.

Nisha Nandhini A, PG Scholar, Department Of Computer Science And Engineering, Karpagam University, Coimbatore, Tamil Nadu, India . (Email: pranisha2530@gmail.com)

Mr. Karthik, Assistant Professor, Department Of Computer Science And Engineering, Karpagam University, Coimbatore, Tamil Nadu, India

Mr.J.Prakash, Assistant Professor, Department Of Computer Science And Engineering, Hindusthan College Of Engineering And Technology, Coimbatore, Tamil Nadu, India. (Email: jeevaprakash86@gmail.com)

Dr.P.Ezhilarasu, Associate Professor, Department Of Computer Science And Engineering, Hindusthan College Of Engineering And Technology, Coimbatore, Tamil Nadu, India.(Email: prof.p.ezhilarasu@gmail.com)

III. KNN

KNN (k- nearest neighbour)[12] is a *non parametric lazy learning* algorithm. That is a pretty concise statement. When you say a technique is non parametric, it means that it does not make any assumptions on the underlying data distribution. This is pretty useful, as in the real world, most of the practical data does not obey the typical theoretical assumptions made (eg gaussian mixtures, linearly separable etc) . Non parametric algorithms like KNN come to the rescue here.

It is also a lazy algorithm. What this means is that it does not use the training data points to do any *generalization*. In other words, there is *no explicit training phase* or it is very minimal. This means the training phase is pretty fast. Lack of generalization means that KNN keeps all the training data. More exactly, all the training data is needed during the testing phase. (Well this is an exaggeration, but not far from truth). This is in contrast to other techniques like SVM where you can discard all non support vectors without any problem. Most of the lazy algorithms – especially KNN – make decision based on the entire training data set (in the best case a subset of them).

The dichotomy is pretty obvious here – There is a nonexistent or minimal training phase but a costly testing phase. The cost is in terms of both time and memory. More time might be needed as in the worst case, all data points might take point in decision. More memory is needed as we need to store all training data.

- *Assumptions in KNN*

Before using KNN, let us revisit some of the assumptions in KNN. KNN assumes that the data is in a *feature space*. More exactly, the data points are in a metric space. The data can be scalars or possibly even multidimensional vectors. Since the points are in feature space, they have a notion of distance – This need not necessarily be Euclidean distance although it is the one commonly used.

Each of the training data consists of a set of vectors and class label associated with each vector. In the simplest case, it will be either + or – (for positive or negative classes). But KNN, can work equally well with arbitrary number of classes.

We are also given a single number "k". This number decides how many neighbors (where neighbors are defined based on the distance metric) influence the classification. This is usually a odd number if the number of classes is 2. If k=1 , then the algorithm is simply called the nearest neighbor algorithm.

- *Pattern recognition with missing data: a review*

The missing values can be denned as the absence of information in instances, which brings harmful consequences to the validity of the subsequent analyzes. The pattern recognition methods are utilized for the applications such as text mining, 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. Pattern recognition is a branch of machine learning that focuses on the recognition of patterns and regularities in data, although it is in some cases considered to be nearly synonymous with machine learning.

- *Imputing missing values: the effect on the accuracy of classification*

Microarray data frequently contain missing values (MVs) because imperfections in data preparation steps (e.g., poor hybridization, chip contamination by dust and scratches) create erroneous and low-quality values, which are usually discarded and referred to as missing. 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.

- *Supervised learning from incomplete data via an EM approach*

In real world the data collected from various organisations consist of incomplete data. The 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. Expectation to maximization is an effective technique that is often used in data analysis to manage missing data. EM imputations are better than mean imputations because they preserve the relationship with other variables, which is vital if you go on to use something like Factor Analysis or Regression. They still underestimate standard error, however, so once again, this approach is only reasonable if the percentage of missing data is very small.

- *Missing data imputation for fuzzy rule-based classification systems*

Many existing industry consist of many missing data. Some of the reason behind them are manual entry of data, operating the machine manually, system error stc.. 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 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.

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

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure.

- The k-means algorithm is an algorithm to cluster *n* objects based on attributes into *k* partitions, where $k < n$.
- It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data.
- It assumes that the object attributes form a vector space.
- Simply speaking k-means clustering is an algorithm to classify or to group the objects based on attributes/features into K number of group.
- K is positive integer number.
- The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

IV. How The K-Mean Clustering Works

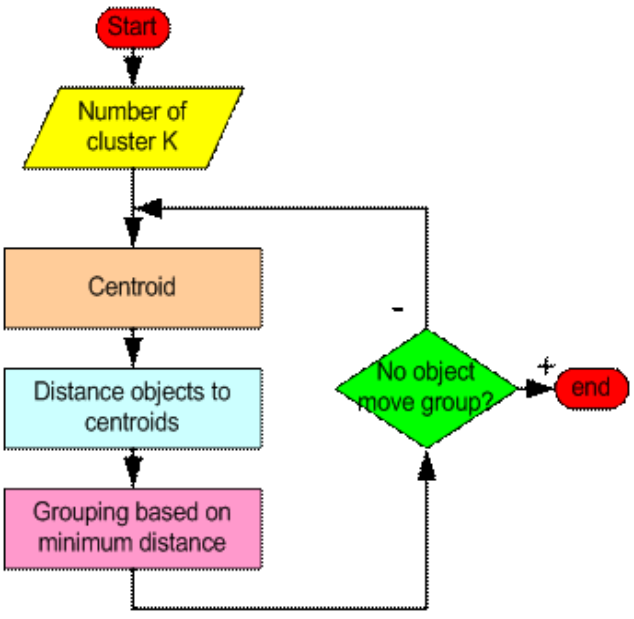

Fig.2. Flow chart for K-Mean clustering

- *Flow chart for K-Mean clustering fig.2 is explained below.*

1. Begin with a decision on the value of K= number of clusters.

2. Put any initial partition that classifies the data into k clusters. You may assign the training samples randomly, or systematically as the following.

1.Take the first k training sample as single- elements clusters.

2.Assign each of the remaining (N-k) training sample to the cluster with the nearest centroid, after each assignment, recomputed the centroid of the gaining cluster.

3. Take each sample sequence and the centoid of each of the clusters. If a sample is not currently in the cluster with the closet centroid , switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.

4. Repeat step 3 until convergence achieved, that is until a pass through the training sample cause no new assignments.

The missing data imputation method based on the most popular techniques in Knowledge Discovery in Databases (KDD), i.e. clustering technique [5]. It combines 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.

- *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.

- *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.

- *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.

- *Neural Network Classifier Based on Dempster-Shafer Theory*

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

- *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.

V. 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.
[11] http://www.thearling.com/text/dmwhite/dmwhite.htm
[12] https://en.wikipedia.org/wiki/Nearest_neighbour_algorithm