International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 9 Issue 1 –JULY 2014.

Cluster Interfaced Objective Function in clustering-based feature selection algorithm for Mining Data

B.Suganya[#1] and M.Pachiyammal[*2]

[#]*PG Scholar, Bharathiyar Arts & Science College for Women*

[*] *Asst., Professor (CSE), Bharathiyar Arts & Science College for Women*

Abstract-Conventional feature selection classifiers work with known and precise data values. In recent data collection methods, appreciable amount of attributes are uncertain. The uncertain attributes, in almost all applications, have more influences on the data set on information classification and feature selectio constructs. Uncertainty needs to be handled properly. Reasons for uncertainty are due to Measurement errors, Quantization errors, Data staleness and multiple repeated measurements. Uncertainty of a data item is represented in terms of multiple values. Usually uncertain data are abstracted by statistical derivatives (eg., mean, standard deviation, median etc.,). Complete information of the data item improves the accuracy of feature selection classifier. In this paper, Proposal work is made to improve the pruning of feature selection classifier algorithm by clustering with distance boundaries and partitioning of uncertain probability distribution values. Clustering techniques increase the speed of feature selection construction and minimize the pruning time to greater extent. Distance boundary clustering technique, works based on the criteria of lower and upper bounds distances of the uncertain attributes values. Partitioning is done with objective function introduced on probability distribution based on the density levels. Objective function introduced evaluates the discrete value of the uncertain data item. Experiments are planned to conduct performance evaluation of heart disease diagnosis and prediction from UCI repository data sets.

Keywords: Objective Function, feature selection, Cluster Interface, Uncertainty.

I. INTRODUCTION

Classification is a traditional crisis in machine learning and data mining. Known a set of training data tuples, every one contains a class label and being denoted by a characteristic vector, the mission is to algorithmically construct a model which predicts hidden test tuple class label based on the tuple's characteristic vector. The well known classification model is the feature selection model. Feature selection are fashionable because they are realistic and simple to recognize. Rules will also be mined from feature selection easily. Numerous algorithms, such as ID3 and C4.5, have been devised for feature selection creation. These algorithms are widely accepted and used in a broad variety of applications such as image recognition, medical diagnosis, and credit rating of loan applicants, scientific tests, fraud detection, and target marketing. Data uncertainty takes place obviously in several applications because of a variety of reasons. We briefly discuss three categories here, 1) Measurement Errors, Data attained from measurements by physical devices are often inaccurate because of measurement errors. 2) Data Staleness, in few applications, data values are continuously varying and evidenced information is constantly stale. Example is location-based tracking system and 3) Repeated Measurements, Perhaps the most familiar source of uncertainty occurs from repetitive measurements. For example, a patient's body temperature could be taken multiple times during a day, an anemometer could record wind speed once every minute.

II. LITERATURE REVIEW

Uncertain data management is one of the important study interests in latest years. Data uncertainty generally categorized into existential uncertainty and value uncertainty. Existential uncertainty happens when it is uncertain either an object or a data tuple occurs. For example, a data tuple in a relational database will be correlated with a probability which denotes the confidence of its occurrence [1]. Probabilistic databases was concerned to semi structured data and XML [2]. Value uncertainty, happens when a tuple is identified to exist, but its values are not identified exactly. A data item with value uncertainty is typically denoted by a PDF over a limited and bounded region of promising values [3].

In [4], the famous k-means clustering algorithm is expanded to the UK-means algorithm for clustering uncertain data. Data uncertainty is typically captured by PDFs that are usually denoted by sets of example values [5]. Extracting uncertain data is hence computationally expensive because of information detonation (sets of example versus single values). To expand performance of UK-means, pruning techniques have been presented [6].

Feature selection classification on uncertain data was discussed for decades in the form of absent values [7]. Absent values emerge when few attribute values do not exist either data collected works or due to data admission errors [8].

Solutions comprise approximating absent values with the mass value or inferring the absent value (either by correct or probabilistic values) using a classifier on the attribute (e.g., ordered attribute tree [11] and probabilistic attribute tree [9]). In C4.5 [10] and probabilistic feature selection [12], missing values in training data are handled by using fractional tuples [13].

Based on the formerly discussed methods, a plain method of "filling in" the absent values might be adopted to handle the absent values [14], taking gain of the capability of managing random PDFs in our method. We take the PDF of the attribute in question over the tuples, where the value is there. The effect is a PDF, which is used as a "guess" distribution of the attribute's value in the absent tuples. Then, we continue with feature selection construction.

III. Performance Evaluation on Cluster Interfaced Objective Function for Feature Selection Classifiers

To explore the potential of achieving superior classification accuracy by considering data uncertainty, we have implemented Cluster Interfaced Objective Function and applied them to heart disease diagnosis and prediction taken from the UCI Machine Learning Repository. The data set is chosen because it contains mostly numerical attributes generated from measurements. For the point of our experiments, classifiers are constructed on the numerical attributes and their "class label" attributes. The attributes of the dataset used are given in the table 1.

Table 1: Pruning Effectiveness

Number Of Entropy Calculation	Pruning Effectiveness (%)	
	UDT-ES	Cluster Interfaced Objective Function
100	69	99
200	68	98
300	60	96
400	64	97
500	67	84
600	62	85
700	54	83

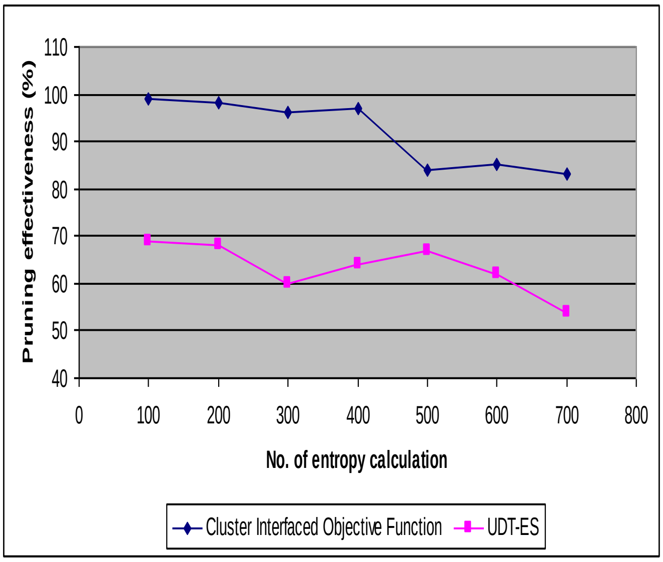

Fig 1 : Pruning Effectiveness

In this section, we study the pruning effectiveness of the Cluster Interfaced Objective Function. Fig. 2 depicts the number of entropy calculations performed by Cluster Interfaced Objective Function and UDT-ES. As we have explained, the computation time of the lower bound of an interval is comparable to that of computing entropy. Therefore, for Cluster Interfaced Objective Function, the number of entropy calculations includes the number of lower bounds computed. The figure 2 illustrates that our pruning techniques introduced are highly effective. Comparing the techniques in resultant graph against that for Cluster Interfaced Objective Function, it is clear that a lot of entropy calculations are avoided by our bounding techniques. By pruning end points, Cluster Interfaced Objective Function minimizes the number of entropy calculations and increasing the pruning efficiency. It thus achieves a pruning effectiveness ranging from 83 % up to as much as 99 %. As entropy calculations control the execution time of Cluster Interfaced Objective Function, such effective pruning techniques significantly reduce the tree construction time.

IV. Conclusion

In this paper we have presented the Cluster Interfaced Objective Function for Feature selection Classifiers for Mining Uncertainty data. Pruning of feature selection classifier algorithm has been improved by clustering with distance boundaries and partitioning of uncertain probability distribution values. Clustering is achieved by Distance boundary clustering technique, based on the criteria of lower and upper bounds distances of the uncertain attributes values. Partitioning and estimating the discrete value of uncertain data is done by Objective function. Relative entropy measure is made on the lower and upper bounded distances on the attribute characteristics related to other certainty attributes in the data set. Experimental results carried out with the metrics as Pruning Effectiveness, number of entropy calculations and Execution time which achieve much better classification accuracy.

V. REFERENCES

[1] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, and Sau Dan Lee, "Feature selectios for Uncertain Data", IEEE Transactions On Knowledge And Data Engineering, vol. 23, no. 1, 2011

[2] Xiaofeng Zhu, Shichao Zhang, Zhi Jin, Zili Zhang, and Zhuoming Xu, "Missing Value Estimation for Mixed-Attribute Data Sets", IEEE Transactions On Knowledge And Data Engineering, vol. 23, no. 1, 2011

[3] E. Hung, L. Getoor, and V.S. Subrahmanian, "Probabilistic Interval XML," ACM Trans. Computational Logic (TOCL), vol. 8, no. 4, 2007.

[4] J. Chen and R. Cheng, "Efficient Evaluation of Imprecise Location-Dependent Queries," Proc. Int'l Conf. Data Eng. (ICDE), pp. 586- 595, Apr. 2007.

[5] M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain Data Mining: An Example in Clustering Location Data," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), pp. 199-204, Apr. 2006.

[6] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J.S. Vitter, "Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 876-887, Aug./Sept. 2004.

[7] R. Cheng, D.V. Kalashnikov, and S. Prabhakar, "Querying Imprecise Data in Moving Object Environments," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1112-1127, Sept. 2004.

[8] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip, "Efficient Clustering of Uncertain Data," Proc. Int'l Conf. Data Mining (ICDM), pp. 436-445, Dec. 2006.

[9] S.D. Lee, B. Kao, and R. Cheng, "Reducing UK-Means to KMeans," Proc. First Workshop Data Mining of Uncertain Data (DUNE), in conjunction with the Seventh IEEE Int'l Conf. Data Mining (ICDM), Oct. 2007.

[10] H.-P. Kriegel and M. Pfeifle, "Density-Based Clustering of Uncertain Data," Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 672-677, Aug. 2005.