International Journal of Information Science 2012, 2(3): 19-22
DOI: 10.5923/j.ijis.20120203.03

Enhanced Active Learning in Developing Highly Interpretable Decision Support System

Mohd Najib Mohd Salleh*, Nazri Mohd Nawi

Faculty of Computer Science and Information Technology, UTHM, Parit Raja, Batu Pahat, Johor, 86400, Malaysia

Abstract Developing highly interpretable decision rules commonly presents significant challenges to decision support system. In previous research work, partial information had provided complex decision in the problem of learning classifiers. The behaviour of some learning algorithm may only be explored by uncertainty analyses. We propose a novel information extraction by utilizing weighted active learning fuzzy measure based on objective function to quantify the goodness of cluster models that comprise prototypes and data partition in decision modelling. By choosing appropriate weights for pre labelled data, the nearest neighbour classifier consistently improves on the original classifier.

Keywords Uncertainty, Decision Support System, Fuzzy Cluster Analysis

1. Introduction

In recent year, the common encountered challenge in decision support system is the occurrence of partially information. This underlying obstacle will return poor and complex decision when the aggregation of all the databases[1]. The behavior of complex decision modeling is coupled with a high degree of uncertainty in estimating the values of many input parameters[8]. From knowledge provider's perspective, the challenge is to gain an expert user's attention in order to assure that the incomplete information is valued[2]. One of the more widely applied in representing uncertainty is fuzzy measure[3,4]. This value would indicate the degree of evidence of the element's membership in the set. Since fuzzy measure can be used to describe imprecise information as membership degree in clustering, therefore a given datasets are divided into partitions based on similarity. We present weighted active learning[5] based on objective function with appropriate fuzzy measure to quantify the goodness of cluster models that comprise prototypes and data partition in decision modelling. In classification tasks, it is generally more important to correctly classify the uncertainty class instances. However in classification problems with partial information, the class examples are more likely to be misclassified. Due to their design principles, most of the machine learning algorithms optimizes the over all classification accuracy hence sacrifice the prediction performance on the partial information. This paper proposes an efficient active learning framework which has high prediction performance to overcome this serious data mining problem.

* Corresponding author:
najib@uthm.edu.my (Mohd Najib Mohd Salleh)
Published online at http://journal.sapub.org/ijis

2. Decision Modeling Construction

The deterministic decision modelling is formulated to assess physiological analysis in planting material selection. All available information is applied to derive the fuzzy maximum estimation which describes the imputation of estimates weight for incomplete information in cluster centres. In decision modelling[9], the attribute depends on its entropy computation among the rest of the attributes, it can be simply formulated the uncertain subset x_i from the interval of element x_o. The entropy can be measured as the item sets are divided into subset T_j

$$\sum_{j=1}^{n} \frac{freq(C_i, T_j)}{T_j} \log_2 \frac{freq(C_i, T_j)}{T_j} \tag{1}$$

In the measured value x_i, the possible subset is uniquely determined by the characteristic function. The measurement of the dispersal range of value x_i is relative to the true value x_o. In early stage, by integration of expert knowledge in planting material, it seem reasonable to calculate the number of samples in the set T_j which belong to the same predicted class C_j., assigned as

$$freq(C_i, T) = \sum_{j=1}^{n} \mu(C_j) \tag{2}$$

When the value of each attribute is assigned to their subset x, $\mu(x)$ represents the degree of available evidence that a given element of X belongs to the subset A. We estimate an attribute x_i provides data in fuzzy set and gives a membership degree of the available evidence. Some data sets can not be separated with clearly defined boundaries due to the

databases which contain incomplete information, but it is possible to sum the weight of data items belong to the nearest predicted class.

3. Proposed Method

We conducted our evaluation on by collecting 2500 examples of planting material in physiological analysis. These records are represented as a table of examples which described by a fixed number of features along with a predicted label denoting its class. In fuzzy decision tree (FDT), fuzzy decision modelling analysis[7,10] refers to all the unique values which some examples may contain incomplete information with discrete set of values or continuous attributes. Figure.1 shows the comparison of complete and incomplete information of physiological analysis during the initial study.

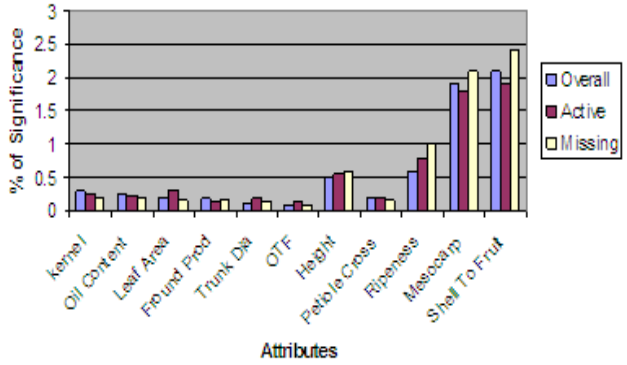

Figure 1. Number of complete and missing attributes in physiological analysis

3.1. Active Learning FDT (ACFDT)

The proposed active learning classifier for fuzzy decision tree (ACFDT) is based on novel idea of clustering input data into sub-group and combine with decision tree classifier. An algorithm for active learning classifier is applied to an initial set L of labeled examples provided by an expert knowledge. Subsequently, sets of M examples are selected in phases from a set of input when the highest weights in local datasets have more influence to objective function.

We define an initial labeled set L as input, an unlabeled set as UL, an inducer as I, a stopping criterion, and an integer M specifying the number of actively selected examples in each phase.

Active Leaning algorithm

1 While stopping criterion not met /* perform next phase: */

2 Apply inducer I to L

3 For each example $\{ x_i \in UL \}$ compute w_i as the effectiveness weight

4 Select a subset S of size M from UL based on w_i

5 Remove S from UL, label examples in S, and add S to L

There are a number of groups within a class into several clusters and we take strong clusters for each class by measuring the highest weight contributes from the input features.

Using the method in Lin[6], the distance between plausible datasets in partial information with local datasets to be minimum numbers of attribute is defined as

$$dist(w, X) = \min_{w' \in \Omega(X)} dist(w, w') \tag{3}$$

Formally, the distance between datasets W, such as weather, fertilizer, land degradation, soil erosion and climate variability able to help in determining the physiological analysis during planting material selection. To take into account the weights of the local datasets as overall distance can be defined:

$$\sum_{k=1}^{n} (dist(w, X) * \mu(X_k)) \tag{4}$$

We describe the objective function in active learning to quantify the goodness of cluster models that comprise prototypes and data partition as

$$l = \sum_{k=1}^{p} \sum_{i=1}^{n} (dist(w, X) * \mu_k(x_i))^m \tag{5}$$

p is the number of specified clusters, k is the number of data points, x_i is the i-th data points, $\mu_i(x_i)$ is a function that returns the membership of x_i in the j-th cluster.

Since fuzzy measure can be used to describe imprecise information as membership degree in fuzzy clustering[11], therefore a given item sets are divided into a set of clusters based on similarity. Fuzzy cluster analysis assigned membership degrees with highest weights in local datasets to deal with data that belong to more than one cluster at the same time; they determine an optimal classification by minimizing an objective function. It can be seen that the characteristic functions constitute between ambiguity and certainty to the closed subsets x_i of the boundary with respect to the set interval. We use more robust method of viewing the compatibility measurement for query x_i takes the weighted average of the predicate truth values. The equation show how the weighted average compatibility index is computed. The average membership approach computes the mean of the entire membership values.

$$x_i = (\sum_{i=1}^{n} (\mu_i(p_i)) \times w_i) / \sum_{i=1}^{n} w_i \tag{6}$$

The degrees of membership to which an item sets belongs to the different clusters are computed from the distances of the data point with additional weights to the cluster centres. The combination of expert knowledge and most relevant ecological information, the membership degree can be calculated to determine some plausible data point lies to centre of the nearest cluster. An iterative algorithm is used to solve the classification problem in objective function based clustering: since the objective function cannot be minimized directly, the nearest cluster and the membership degrees are alternately optimized.

4. Experiments and Analysis

We experimented on incomplete information in ecological system impacts on classification evaluation of planting

material behaviour in physiological analysis. In figure 2, the graph shows the observed data are more likely to satisfy with additional information. The experiment combines planting material with expert knowledge and ecological information to generate an ROC curve. It can be seen that, the highest collective weights is tuned to minimize the number of attributes chosen during the classification. Therefore, in most of the data sets belonging to the various classes exactly match the results of physiological analysis. The classical decision tree and fuzzy decision tree approach reports the smallest variance on the figure, while active learning fuzzy decision tree showed significant improvements over random decision tree methods when apply weights.

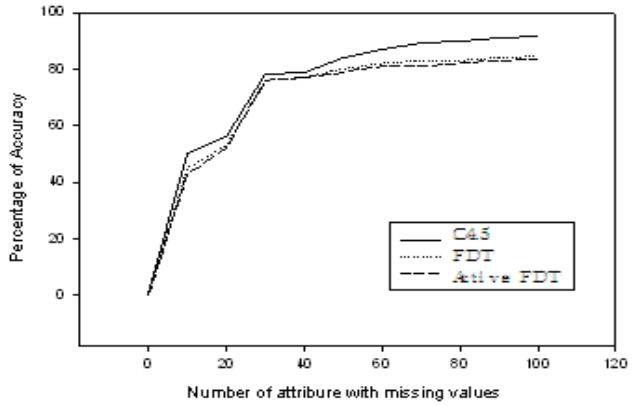

Figure 2. The accuracy of predicted class using C4.5, FDT and Active Learning FDT

In this experiment, we examine gradually the approaches of decision tree with selected complete and incomplete examples, ranked features and missing records of real physiological traits. The result, in Table.1 shows that the features in missing values compare to the others complete examples. By adding additional features of ecological information to physiological trait, it shows less correlation between each feature in missing values and the others.

Table 1. Summary of comparisons of agrees and disagree classification based on decision tree algorithms

	Training		Testing	
Modeling Decision with Fuzzy Set	Agree	Disagree	Agree	Disagree
	83.37%	16.63%	81.09%	18.91%
Modeling Decision with incomplete information	Correct	Wrong	Correct	Wrong
	78.82%	21.18%	76.68%	23.32%

The possible clusters rearrange their structure and rules are generated mostly come from combination of planting material and ecological information. As a result, the selected proper weighted fuzzy measures in decision tree construction provide less and simple rules in Figure 3. It removes some irrelevant features during the selection of subset of training data.

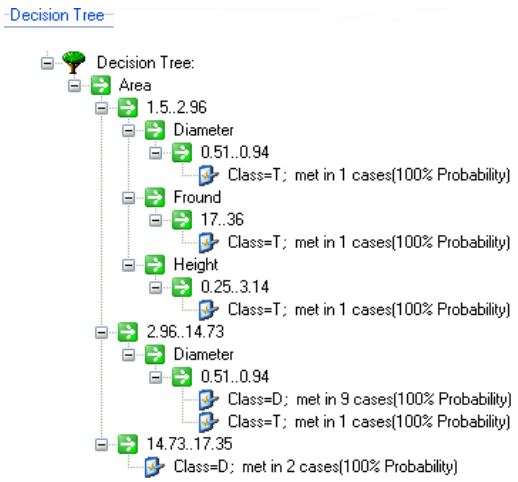

Figure 3. The production rules by proposed Active Learning FDT method

5. Discussion

In this paper, we proposed weighted active learning on fuzzy decision tree, has focused on domains with relatively high levels of unknown values and small testing set. The method partitions a subspace into non-overlapping cluster by utilizing fuzzy measure collective weights for small samples and produces better result. We have measured the degree of uncertainty using membership function and the proposed method is near to the true values of the uncertainty and the relative errors of the estimation are very small. However, this method still produces fairly large error after pruning with more missing values. This shows that the proposed method can only be applied to several conditions. The uncertainty estimation of measured values can be directly obtained using fuzzy representation and it is no longer necessary to estimate the standard deviation by mean value. For systems with small samples and unknown distributions, the proposed method is more suitable. When the training set is partitioned, ignoring cases with unknown values of the tested attribute leads to very inferior performance. During classification, attempting to determine the most likely outcome of a test works well in some domains, but poorly in others. Combining all possible outcomes is more resilient, giving better overall classification accuracy in these domains.

In the last experiment, we tested active learning FDT using real data to produce decision rules. The active leaning FDT algorithm can exploit class specifically occurring missing values for classifying the artificially corrupted dataset. The data contains three different classes with Dura and Psisifera gene type.

6. Conclusions

The uncertainty is not only due to the lack of precision in measured features, but is often present in the model itself

since the available features may not sufficient to provide a complete model to the system. We have focused on fuzzy measure with collective weight to model uncertainty in decision support system. In this study, a special treatment of uncertainty is presented using fuzzy representation and clustering analysis approach in constructing the decision model. The result of the study shows that uncertainty is reduced and several plausible attributes should be considered during classification process. This formalization allows us to the better understanding and flexibility for selecting planting material in acceptance of the classification process.

ACKNOWLEDGEMENTS

The author would like to thank Universiti Tun Hussein Onn Malaysia for funding the research grant Vot 0705.

REFERENCES

[1] Z.S.Chao,Q.Z.Xing,C.X.Ling and S.S.Li, "Missing is Useful: Missing Values in Cost-Sensitive Decision Trees", IEEE transactions on knowledge and data engineering, vol. 17, no. 12, December 2005.

[2] W.B.Rouse,"Need to know - Information, Knowledge and Decision Maker", IEEE Transaction On Systems, Man and Cybernatics-Part C: Applications and Reviews, Vol 32, No 4, November 2002.

[3] J.G .Klir. and A.T. Folger. "Fuzzy Sets, Uncertainty and Information", Prentice Hall, 1988.

[4] M.X.Wang and C. Borgelt, "Information Measures in Fuzzy Decision Tree", Fuzzy Systems, Proceedings and IEEE International Conference,25-29, July 2004.

[5] Monteleoni.C, Kaariainen. M,"Practical Online Active Learning for Classification", IEEE.

[6] J. Lin "Framework for Dealing with Confliction Information and Application". Ph.D Thesis, Dept. of Computer Science, University of Toronto. 1995.

[7] F. Mendonca, S.M Vieira, J.M Sousa, "Decision Tree search Methods in Fuzzy Modeling and Classification", International Journal of Approximate Reasoning, 2007.

[8] C.R.Michelini and B.G. Rossi "Measurement uncertainty: a probabilistic theory for intensive entities Measurement", 15 143–57, 1995.

[9] J.Quinlan, "Introduction to Decision Tree", Morgan Kaufman 1993.

[10] M.Tokumaru, N.Muranaka, "Product-Impression Analysis Using Fuzzy C4.5 Decision Tree", Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol13 No.6, 2009.

[11] T.Heiko, D.Christian, K.Rudolf, "Different Approaches to Fuzzy Clustering of Incomplete Datasets", International Journal of Approximate Reasoning 35 239-249. 2004.

[12] H. Timm, C. Doring, R. Kruse, Fuzzy cluster analysis of partially missing datasets, in: Proceedings of the European Symposium on Intelligent Technologies, Hybid Systems and Their Implementation on Smart Adaptive Systems (EUNITE 2002), Albufeira, Portugal, 2002.