IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661, ISBN: 2278-8727 Volume 5, Issue 3 (Sep-Oct. 2012), PP 45-51
www.iosrjournals.org

Classification By Clustering Based On Adjusted Cluster

Priyamvada ojha[1], Pradeep sharma [2]
[1,2] (Computer Science Engineering , IEC college of engineering ,Greater Noida / Mahamaya Technical University, INDIA)

Abstract: *Currently cluster analysis techniques are used mainly to aggregate objects into groups according to similarity measures. Whether the number of groups is pre-defined (supervised clustering) or not (unsupervised clustering), clustering techniques do not provide decision rules or a decision tree for the associations that are implemented. The current study proposes and evaluates a new technique to define decision tree based on cluster analysis. The proposed model was applied and tested on two large datasets of real life HR classification problems. The results of the model were compared to results obtained by conventional decision trees. It was found that the decision rules obtained by the model are at least as good as those obtained by conventional decision trees. In some cases the model yields better results than decision trees. In addition, a new measure is developed to help fine-tune the clustering model to achieve better and more accurate results.*

Keywords:*Classification,Classifier,Cluster analysis ,Decision trees decision rule,Imbalanced data.*

I. Introduction:

Currently, decision trees provide useful solutions for many classification problems related to large datasets that often containmissing values or errors (Aitkenhead, 2008). Decision trees act like a ‘‘white box’’ which gives the user a good understanding and easy interpretation of the results. Typically, decision trees are used to resolve classification problems by constructing rules for assigning objects to classes (Jamain& Hand, 2008). Despite the strengths of decision trees, generating a significant decision tree model can be impeded by the nature of the dataset. Classification trees can be unstable and sensitive to small variations in the data, such as those caused by randomization (Adhanom, 2009), making it impossible to obtain clear classification rules. This weakness can only surface in practical uses of decision trees and for this reasonis rarely discussed in the academic literature. Nevertheless there is a genuine need for a method that can handle classification problems in particular in those cases where decision trees fail to provide a meaningful decision rule.

The current study proposes a decision tree construction method based on a preliminary analysis using cluster analysis techniques. The method is dubbed ‘‘classification by clustering’’ (CbC) because the decision trees/rules are based on adjusted cluster analysis. Conventional decision trees are defined along a recursive partitioning in whichthe choice to split attributes involves picking the attribute that will partition the original sample into sub-samples that are as homogenous as possible in relation to the class variable (Adhanom, 2009). The proposed model presents a new approach:instead of trying to find statistical associations between the attributesand the class variable it is based on similarities, the core conceptin cluster analysis techniques.This model is tested against decision tree algorithms on tworeal life HR classification problems. The findings show that the resultsare as good as and in some cases even better than results obtainedby conventional decision trees and can yield a meaningfulclassification even in those cases where the decision trees failedto provide one. In addition, we propose two new measures basedon the Mean Square Error (MSE). One measure is used to assessthe model’s results in cases where conventional measures (lift, precision,recall, etc.) are not significant and therefore cannot be used.The other measure is used to determine the weight of each attributein the fine-tuning stage.

The model and the measures can be used on any classificationproblem that has a binary target variable. CbC thus expands andenriches the available set of tools for such problems, and contributesto resolving problematic datasets that cannot be classifiedby conventional ‘‘statistical’’ decision trees.

II. Background review:

2.1. Literature review

Data mining (DM) is rapidly becoming a front runner in the academic and commercial area of managing and utilizing company data resources (Agrawal, 1999). The objective of DM is to detect, interpret and predict qualitative and quantitative patterns in data, leading to an incremental level of information and knowledge A wide variety of models and algorithms are employed, from statistics, artificial intelligence, neural nets and databases to machinelearning. Thisarticle discusses classification problems, which is a specific caseprediction problems. In these problems the objective is to assigna new object to its correct class (that is, the correct Y category) onthe basis of its observed X values Usually, classification is based on the statistical probability toobtain each one of the possible values of the target attribute. However,we propose a novel approach, based on clustering principles,in which the classification is based on similarity-association. Clusteringrefers to decomposing or partitioning a dataset into groupsso that the points in one group are similar to each other and areas different as possible from the points in other groups . Clustering models do not use target attributes; ratherthey partition the dataset by using similarity measures . Much of the researchin the area of clustering attempts to create better algorithmsand better suited datasets for the clustering process. Forexample presentedmethods to choose entities to be used as centers in ‘‘centerapproach’’ algorithms such as the k-means algorithm, in order toimprove their performance. Others, such as , described representation methods to support improvedsimilarity functions, and fuzzy data elements.Studies such as applied theoretical concepts to specific real life problemsand datasets that tend to experience much more noise and uncertainty than synthetic datasets.

It is well-known that cluster analysis involves subjectivity asdoes any similarity-distance measure (due to expert assignmentof attribute weight). In addition, the same dataset is often partitionedin different ways by different applications . Different clustering models at times provide verydifferent results, and there is no way of knowing which is the rightone or the best . Nevertheless, clustering models have a crucial strength in that they always providea result, whereas in real life classification problems, decisiontrees can fail to do so (generating instead a small and insignificanttree with poor assessment measures. In such cases the researcherencounters a dead end because it is impossible to further analyzethe dataset.

So far, most studies on classification problems have only usedconventional models, mainly decision trees. In an overall assessment of more than 5000 classification problems known in the literaturepresented by Jamain and Hand, none was handled byusing clustering algorithms (Jamain& Hand, 2008). Currently,there is still a clear division between clustering methods and decisiontrees which are still considered the main method to handle classification problems.

2.2. Overview of classification techniques:

The following section schematically describes the classificationtechniques used in this study.

_ Classification and Regression tree (C&R tree) – a tree-based classificationmethod which uses recursive partitioning to split the training records into segments with similar output field values.

The C&R tree starts by examining the input fields to find thebest split, as assessed by the reduction in an impurity index thatresults from the split. The split defines two subgroups, each ofwhich is subsequently split into two more subgroups, and soon, until one of the stopping criteria is triggered. All splits are binary (SPSS, 2003).

- CHi-squared Automatic Interaction Detection Tree (CHAID Tree) - a method for building decision trees that uses chi-square statistics to identify optimal splits. CHAID first examines the cross-tabulations between each of the predictor variables and the outcome, and tests for significance using a chi-square independencetest. If more than one of these relations is statistically significant, CHAID will select the predictor that is the most significant, i.e., that has the smallest p-value (SPSS, 2003).
- K-means – a clustering algorithm (MacQueen, 1967) which is available in many statistical and data mining tools. The algorithm divides the dataset into a pre-determined number of clusters and contains the following steps: (i) choose k –cluster centers randomly from the points (patterns) in the dataset. (ii) Assign each pattern to the closest cluster center. (iii) Re-compute the cluster centers using the current cluster membership. (iv) If a convergence criterion is not met, go to step 2. Typical convergence criteria are: no (or minimal) reassignment of patterns to new clusters, or minimal decrease in squared error. Several variations of the k-means algorithm have been reported in the literature (Jain, Murty, & Flynn, 1999).
- Two Step – a clustering algorithm consists of two passes over the dataset. The first pass divides the dataset into a coarse set of sub-clusters, while the second pass groups the sub-clusters into the desired number of clusters. The desired number of clusters can be determined automatically, or it can be a pre determined fixed number of clusters (Gelbard et al., 2007).

2.3. Overview of evaluation measures:

The following section schematically describes the evaluation measures used in this study.

- Lift – expresses the improvement of the prediction achieved bythe model, compared to the existing state. The maximum possiblelift is calculated as 1/R where R is the total response rate ofthe population (SAS, 2005).
- Precision – the number of true positives (i.e., the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e., the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class)
- Recall – the number of true positives divided by the total number of elements that actually belong to the positive class (i.e., the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

- F-score – a measure of a model's accuracy. This considers both the precision and the recall of the model to compute the score. The F-score can be interpreted as a weightedaverage of the precision and recall, where an F-score reaches its best value at 1and worst score at 0.
- Root Mean Square Error (RMSE) – this is one of the most commonmeasures in statistics. It is usually used to assess the fitof a model's results to actual results by summing the squaredresiduals (differences between the model value and the actualvalue) dividing the result by the number of observations andthen root calculation. RMSE is a minimal measure; i.e., its bestpossible value is zero. When RMSE equals zero it means themodel perfectly describes the actual values of observations.

III. The proposed model – classification by clustering:

The current study proposes a new model to define a decision tree-like classifier, based on adjusted cluster analysis classification called classification by clustering (CbC). The model is in fact a methodology for decision tree definition based on clustering algorithms.

The main advantage of this model is that it always provides a meaningful decision rule, unlike decision trees that sometimes fail to provide rules that the researcher can actually use.

Like all classification methods, the classification by clustering model (CbC) also uses the methodology of training and test sets.

It is implemented in a machine-learning process composed of six steps, as follows:

Step 1: Choose the target attribute – since it is a classification model, a target attribute is essential. The target attribute must be categorical.

Step 2: Run a clustering algorithm on the dataset – any clustering algorithm can be used. If the algorithm demands parametersfrom the researcher (such as the desired number of clusters) it is recommended to run the algorithm several times using different parameters to find the most parsimonious

Step 3: Calculate the target attribute distribution for each cluster – each of the clusters contains part of the entities in the dataset. This group of entities has its own distribution of the target attribute. If the target attribute is binary, this distribution is called the response rate of the cluster.

Step 4: Set a threshold – the calculated distribution of the target attribute in each of the clusters is actually the probability for an entity with similar attributes to have each of the possible values of the target attribute. Once a threshold is set, all the entities in each group are classified with respect to the appropriate value of the target attribute. For example, if the target attribute is binary and the threshold is 50%, entities of clusters with a response rate
above 50% will be classified as Y and entities of clusters with a response rate below 50% will be classified as N.

Step 5: Fine-tune the results – since clustering models are devised without using a target attribute and do not have a built-in validation process they are often inferior to conventional models. To overcome this problem the results of the clustering algorithm need to be fine-tuned. This is done by giving extra weight to some of the more important attributes
in the dataset. By doing so, it is possible to create clusterswith a stronger correlation to the target attribute. In thelast part of this paper a new measure for the fine tuningprocess is detailed described.

Step 6: Test the results – run the results of the clustering algorithm on a ''fresh'' set of data (test data) and classify the entities accordingly. Because the target attribute of the test data is known, it is possible to assess the results by conventional measures (precision, recall, etc.) or specific measures developed especially for special cases such as explained below.

The output of these six steps is a decision tree-like classifier based on cluster analysis which can be implemented for various classification problems.

For commercial purposes, a good classifier is one that is capable of dividing the population into groups with both significant sizes and response rate, where the distribution of the response rate significantly differs from the response rate of the entire population. In
cases where a resulting class is very small it is not sufficient even if it has a high/low response rate with respect to the entire population.If the class is too small, it is ineffective and will be neglected in real life problems.

For this reason, we incorporate an additional measure for the training and evaluation stages, a Weighted Group Score index (WGS), which is based on the common Mean Square Error (MSE) calculation, with two modifications. The first is that unlike MSE, this measure is maximal i.e., large values of the measure are better than small values. The second relates to weights given to each residual based on the size of the group.

In cases in which it is impossible to draw out significant groupclasses by using conventional decision trees, it is also impossible to use conventional measures (such as lift, precision, recall, etc.) because of their poor and meaningless output in such cases. On the other hand, the WGS measure remains meaningful and can yield proper associations.

The WGS measure is defined by the following formula:

$$\mathrm{WGS} = \frac{\sum_{I}(R_I - R)^2 . N_I}{N}$$

whereiis the group index, R is the response rate in entire population(between 0 and 100), Riis the response rate in group i(between0 and 100), N is the total number of entities, Ni is the number ofentities in group i..

It is clear that this measure generates a high value when thereare groups with different response rates (compared to the generalresponse rate of the population) and large sizes, so it guaranteesthat models that find large groups with a high/low response ratewill be ranked high.

IV. Research method:

4.1. The datasets:

Two datasets were used to test and evaluate the proposed model. Since the model is intended to support real life classificationproblems, both were large real-life datasets. The classification by clustering model was tested and evaluated in comparison to conventional decision tree models. For this purpose various models were run and then compared using a set of measures. The datasets were obtained from a large international company that recruits hundreds of new employees each year from thousands of potential candidates. Because of privacy issues, the actual data items and the attributes are masked. However, both datasets are available in their ‘‘encrypted’’ form through our faculty website. Before recruiting an employee, the company uses a sorting process which enables it to collect relevant information about the candidates. The datasets contain a variety of attributes about each candidate, most of them categorical (ordinal or nominal), but some of which are binary. The target variables of the datasets are binary; i.e., they are assigned a value of either Y/N.

4.1.1. **Dataset 1 – preliminary evaluation ofcandidates-**The sorting process the company uses takes place in severalstages. The first dataset contains data collected during the initialstage of the sorting process. At the end of the first stage, a greatdeal of data has been collected on each candidate. The goal at this early stage is to find candidates who are likely to be dropped beforeor at the end of the initial sorting process (be rejected). Therationale for identifying these candidates early is that the subsequentparts of the sorting process takes time and cost the companya considerable amount of money. The underlying assumption isthat early detection of these candidates can save resources withoutnegatively affecting the sorting process because it will help thecompany avoid spending time and money on unsuitable candidates.The dataset contained data collected using testing tools suchas quantitative and qualitative exams, personal interviews andquestionnaires. The attributes were divided into four groups: (i)scores on six levels of knowledge and education tests, (ii) scoreson three psychological personality tests, (iii) scores on threebehavioral tests, (iv) three other measures provided additional general information about the candidate. The dataset contained data on the candidates processed by the company in the years 2001–2003 for a total of 19,719 records. A target attribute with the value Y indicates a candidate who was dropped during or at the end of the initial sorting process.

The drop rate for 2001–2003 was 44%, 48%, and 55%, respectively. The data were divided into two parts:

- Training – the data used to train and validate the various models. Contained the years 2001–2002. Total number of records: 14,093.
- Testing – the data used to test the models. Contained the year 2003. Total number of records: 5626.

4.2. The classification method

The ‘‘preliminary evaluation’’ dataset (dataset 1) fits the definition of a classification dataset that can be classified using common decision tree models. The ‘‘candidates’ training success’’ dataset (dataset 2) is an example of a classification dataset that cannot be classified using common decision tree models because it is impossible to build a decision tree which enables a significant classification.

The effectiveness of the models, on both datasets, should mirror their effectiveness on other daily classification problems.

All the classification models were built and executed using Clementine 10.1 data mining software by SPSS.

The classification by clustering, for the first problem/dataset, followed the following steps:

1. Divide the dataset into two groups; use the first to build the models and the second to test them (i.e., training and test sets).
2. Create decision trees – 2 decision trees were created (C&R tree and CHAID).
3. Set a threshold – the threshold is the minimum response probabilityneeded to classify the entity target attribute as Y.
4. Create the classification by clustering model – here, the clustering algorithms employed were K-means and Two Step (since a literature review showed them to be superior to other cluster clustering algorithms). Since the K-means algorithm demands a predetermined number of clusters, various cluster numbers were tested.
5. Fine-tune the clustering models – the fine-tuning process compensates for the fact that clustering algorithms do not use a target attribute.
6. Run each of the models on the test group – classify each entity as regards the predicted target attribute according to each of the models and the previously defined threshold.
7. Compare the models with precision and recall measures

.

The F-score measure (which is a weighted average) was not used because both recall and precision have a meaning of their own and it is essential to analyze them separately. Instead, the recall and precision values were presented graphically and an efficiency curve was generated.

Since it is impossible to build a high-quality classifier for the second problem/dataset because a decision tree which enables a significant classification cannot be constructed, two of the steps were slightly different than the above and the adjusted cluster analysis classification was defined as follows:

- Run each of the models on the test group. Because it is impossible to classify each entity to the predicted target attribute significantly, the entities in the test data were divided into groups according to steps 1–7. The actual response rateof each group and the quality of the division were thenassessed.
- Compare the models using WGS (see Section 3).

:

V. Model evaluation:

5.1. Classification of dataset 1: preliminary evaluation of candidates

5.1.1. Classification by decision trees Two decision trees were built based on the dataset, a C&R tree and CHAID. Both of them are well known and have been proven to be effective in classification problems. The dataset was divided into train and test sets at a ratio of 70:30, respectively.

Because the drop percentage in the population is about 50%, the maximum possible lift is about 2. The lift achieved by both models was close to the possible maximum so it is clear that conventional models are appropriate for this problem and can help identify the high risk entities with a high level of certainty.

Setting the threshold – in order to classify the high risk population it is essential to set a threshold. An overly low threshold might create an overly sensitive model where the percent of false positive entities would be very high. On the other hand, an overly high threshold might create a non-differentiating model where the percent of false negative would be very high.

It was decided to set a threshold of 60% in order to increase precision. This decreases the percent of false positive entities with a model which is sensitive enough to identify a sufficient percent of drop entities.

5.1.2. Classification by clustering

The model followed the six steps described above.

Step 1: As mentioned above, the target attribute is binary (Y if the candidate has dropped and N otherwise).

Step 2: The model was implemented using two clustering algorithms: (i) K-Means (from 6 to 11 clusters), (ii) Two Step. The clustering models were built without using the target attribute and without dividing the building data into train .

Step 3: The actual drop rate (the percent of entities with target attribute Y) was calculated for each group.

Step 4: In order to compare the model to the decision trees it was decided to use the same threshold; i.e., every entity belonging to a group with drop rate higher than 60% was assigned to the high risk population.

Table 1 shows the precision and recall comparison based on the validation data (for decision trees) and building data (for clustering models). The results of the Two Step model were inferior so it was decided not to analyze them any further. As seen, the decision trees provide good results because the measures are relatively high. It is clear that the precision of the clustering models is close to the decision trees.However, the recall of the clustering models is lower.

Step 5: The fine-tuning process was done in a trial and error mode. The idea was to give a double or triple weight to various attributes until a satisfactory result was obtained. In the future, more sophisticated algorithms can be used to determine the optimal weight for each attribute, and probably achieve even better results. The model chosen for the fine tuning process was k-means 10 because it providedbetter results than most of the other clustering models. Nevertheless, other models could have been selected. Except for the original model (baseline), nine other combinations were tested. Each combination represents a different set of weights for the attributes in the dataset.

As shown in Table 1, the main problem of the clustering models is low recall measures compared to the decision trees. Therefore, the combinations chosen to be further analyzed were the ones that increased recall.

Step 6: The results were tested and compared to the decision trees based on the test data, as shown in the next paragraph.

5.1.3. Evaluation of the results

The models were tested on a new and unfamiliar dataset (year 2003) i.e., a dataset which was not used to build or validate themodels and the results were compared. Table 2 shows the results for precision and recall data for both models. Maximal values on both measures indicate the better model.

Fig 1 shows the measures graphically; it is clear that there is atrade-off between the two measures. To maximize both measures, the efficiency curve is convex to the origin.

The efficient models are clustering (model 1) and decision tree (model 4). The clustering models achieved higher recall compared to decision trees but their precision was lower. One of the main advantages of the clustering models is their sensitivity; i.e., their ability to identify a larger number of risk candidates (and therefore obtain better recall results). In fact, using these models reduces the likelihood (compared to decision trees) of assigning candidate who were dropped (false negatives). It is also clear that the fine-tuning

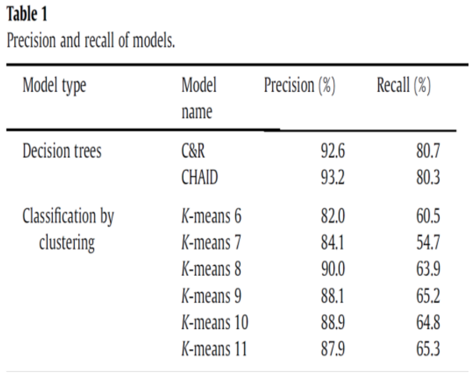

Table 1
Precision and recall of models.

Model type	Model name	Precision (%)	Recall (%)
Decision trees	C&R	92.6	80.7
	CHAID	93.2	80.3
Classification by clustering	*K*-means 6	82.0	60.5
	K-means 7	84.1	54.7
	K-means 8	90.0	63.9
	K-means 9	88.1	65.2
	K-means 10	88.9	64.8
	K-means 11	87.9	65.3

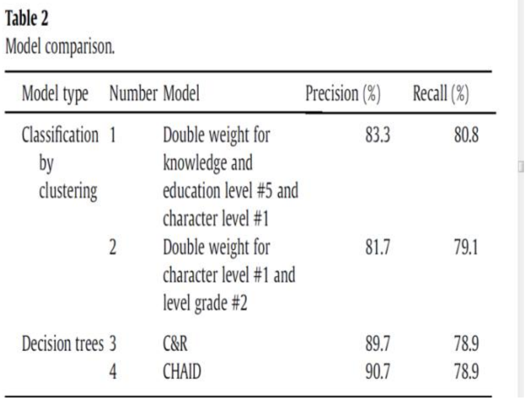

Table 2
Model comparison.

Model type	Number	Model	Precision (%)	Recall (%)
Classification by clustering	1	Double weight for knowledge and education level #5 and character level #1	83.3	80.8
	2	Double weight for character level #1 and level grade #2	81.7	79.1
Decision trees	3	C&R	89.7	78.9
	4	CHAID	90.7	78.9

The models were also compared in terms of thresholds (50% insteadof 60%). The results (as shown in Appendix D) were similar and again prove that the clustering models provide results that are as good as decision trees.

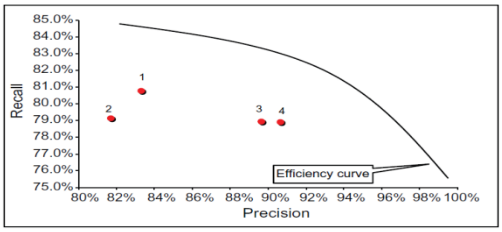

Fig. 1. Graphic display of the model comparison.

VI. Summary and conclusions:

Currently, the use of clustering models is not as widespread as decision trees. This is primarily because clustering models are usually limited to problems of population division. The current research proposed a new method to define a decision tree-like based on adjusted cluster analysis that classifies by clustering.

The model was tested and compared to conventional decision trees on two real life datasets. Because the model was designed to handle real life problems it was essential to test it on real life datasets. The results show that:

- Using the classification by clustering method enables the researcher to obtain classification results that are at least as good as (and in some cases even better) than the results provided by decision trees on both ‘‘good’’ datasets and ‘‘bad’’ datasets.
- The classification by clustering method provides useful and meaningful results even if the dataset is ‘‘bad’’ and conventional decision trees are ineffective. By implementing the classification by clustering method it was possible to build a useful model that divided the population into groups with different response rates than the population average, and significant sizes. The clustering models produced results which were about 20% better than conventional decision trees.

References:

[1] Adhanom, T. (2009). Classification trees. Weka Docs website. <http://wekadocs.com/node/2>Retrieved 08.01.10.

[2] Adriaans, P., &Zantinge, D. (1996). Data mining. Reading, MA: Addison-Wesley.

[3] Agrawal, R. (1999). Data mining: Crossing the chasm. In Proceedings of ACM SIGKDDconference on knowledge discovery and data mining, San Diego, August 15–18,1999.

[4] Aitkenhead, M. J. (2008). A co-evolving decision tree classification method. ExpertSystems with Applications, 34(1), 18–25.

[5] Chung, H. M., & Gray, P. (1999).Data mining.Journal of Management InformationSystems, 16(1), 11–16.

[6] Duda, R. O., & Hurt, P. E. (1973). Pattern classification and scene analysis. NY, USA:John Wiley & Sons.

[7] Elder, J. F., &Pregibon, D. (1996). A statistical perspective on knowledge discovery indatabases. In U. Fayyad, G. Piatetsky-Shapiro,.

[8] Smyth, & R. Uthurusamy (Eds.),Advances in knowledge discovery and data mining (pp. 83–113). AAAI, MIT Press.

[9] Erman, J., Arlitt, M., &Mahanti, A. (2006). Traffic classification using clusteringalgorithms. ACM SIGCOMM, 2006, 281–286.

[10] Estivill-Castro, V., & Yang, J. (2004). Fast and robust general purpose clusteringalgorithms. Data Mining and Knowledge Discovery, 8(1), 127–150.