Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 7, July 2013, pg.365 – 370

RESEARCH ARTICLE

STUDY ON DECISION TREE COMPETENT DATA CLASSIFICATION

A.Vanitha[1], S.Niraimathi[2]

[1]Department of Computer Science (Aided) & NGM College, India

[2]Department & Computer Science & NGM College, India

[1] vanitha24.anandh@gmail.com; [2] niraisenthil@hotmail.com

Abstract— ***Data mining is a process where intelligent methods are applied in order to extract data patterns. This is used in cases of discovering patterns and trends among large datasets. Data classification involves categorization of data into different category according to protocols. They are many classification algorithms available and among the decision tree is the most commonly used method. Classification of data objects based on a predefined knowledge of objects is a data mining. This paper discussed about classification and classifying the kind of structure from a data set according to the behavior.***

Keywords— Data mining; Decision Tree Classification; K-Means Algorithm

I. INTRODUCTION

Data mining is the extraction of implicit, previously unknown and rotationally useful information from data. It is extraction of large database into useful data or information and that information is called knowledge. Data mining is always inserted in techniques for finding and describing structural patterns in data as a tool for helping the data and to make predictions. Data mining consists of five major elements. First to extract, transform, and load transaction data onto the data warehouse system. Second, to store and manage the data in a multidimensional database system. Third, to provide data access to business analysis and IT professional. Fourth, to analyse the data by application software. Fifth, to present the data in a useful format, such as a graph or table. Many data mining techniques are closely related to some of machine learning. Others are related to techniques that have been developed in statistics sometimes called exploratory data analysis. We select clustering k-means algorithm to improve the training phases of classification. Learning classification methods in data mining can be classified into three basic types: supervised, unsupervised and reinforced.

A. Supervised Learning

The class labels of each training tuple are provided [1]. Supervised learning requires that the target variable is well defined and that a sufficient number of its values are given. The method is usually fast and accurate.

B. Unsupervised Learning

The class labels of each training tuple are not known. In unsupervised learning situations all variables are treated in the same way, there is no distinction between explanatory and dependent variables. For unsupervised learning typically either the target variable is unknown or has only been recorded for too small a number of cases.

C. Reinforced Learning

Reinforcement learning is defined not by characterizing learning methods, but by characterizing a learning problem. Reinforcement learning is different from supervised learning, Reinforcement learning is defined not by characterizing learning methods, but by characterizing a learning problem. Any method that is well suited to solving that problem, we consider to be a reinforcement learning method.

II. RELATED WORK

Classification is the processing of finding a set of models or function which describe and distinguish data classes or concepts. The derived model is based on the analysis of a set of training data. The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can be easily converted to classification rules. Decision trees can handle high dimensional data. Decision tree is a mapping from observations about an item to conclusions about its target value [2,3,4,5,6] Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of decision trees is as a descriptive means for calculating conditional probabilities. A decision tree (or tree diagram) is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility [7].

A. Available Classification Techniques

There are so many techniques are available for data classification. In this paper we take into consider only four well known techniques:

1) *Neural Networks*: includes the most popular architecture: a single hidden layer Preceptor with optional short cut connections. We restrict the so-defined model class (which also includes the linear model) to architectures with at least one hidden unit and no short cut connections to prevent a “fall-back” to LDA [8].
2) *Support Vector Machines (SVM):* are used for support vector classification with linear kernel and non-linear kernel functions.
3) *KNN (K- Means):* classifier uses the training data to assign new features to the class determined by the nearest data point. It uses the Euclidean distance measure (not suitable for categorical predictors) to it.
4) *Decision Tree* tries to find an optimal partitioning of the space of possible observations, mainly by the means of subsequent recursive splits.

B. Algorithm Selection

Once preliminary testing is judged to be satisfactory, the classifier is available for routine use. The classifier’s evaluation is most often based on prediction on accuracy. There are at least three techniques which are used to calculate a classifier’s accuracy. One technique is to split the training set by using two-thirds for training and the other third for estimating performance. In another technique, known as cross-validation, the training set is divided into mutually exclusive and equal-sized subsets and for each subset the classifier is trained on the union of all the other subsets. The average of the error rate of each subset is therefore an estimate of the error rate of the classifier. Leave-one-out validation is a special case of cross validation. All test subsets consist of a single instance. This type of validation is, of course, more expensive computationally, but useful when the most accurate estimate of a classifier’s error rates required. Training a standard decision tree leads to a quadratic optimization problem with bound constraints and one linear equality constraints. Training support vector machines involves a huge optimization problem and many specially designed algorithms have been proposed. We used an algorithm called “Decision Tree Induction” that accelerates the training process by exploiting the distributional properties of the training data, that is, the natural clustering of the training data and the overall layout of these clusters relative to the decision boundary of support vector machines.

1) *Sub Process:*

A fast training algorithm called Decision Tree Induction whose idea is to speed up the training process by reducing the number of training data. This is accomplished by partitioning the training data into pair-wise disjoint clusters, each of which consists of either only support vectors or only non-support vectors, and replacing the cluster containing only non-support vectors by a representative. In order to identify the cluster that contains only non-support vectors, the training data is first partitioned into several pair-wise disjoint clusters and an initial support vector machine is trained using the representatives of these clusters [9].

III. K -MEANS ALGORITHM

K-means is the simplest and most popular classical classification and clustering method that is easy to implement. The method is called k-means since each of the K clusters is represented by the mean of the objects within it. It is also called the centroid method since at each step the centroid point of each cluster is assumed to be known and each of the remaining points are allocated to the cluster whose centroid is closest to it. Once this allocation is completed, the centroid of the clusters are recomputed using simple means and the process of allocating points to each cluster is repeated until there is no change in the clusters [10]. The K-means algorithm proceeds as follows. First, it randomly selects k of the objects, each of which initially represents a center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster. It then computes the new mean for each cluster. This process iterates until the criterion function converges. The K-Means algorithm is well known for its efficiency in clustering large data sets. K-means clustering is a method of cluster analysis which aims to partition *n* samples of dataset into *k* clusters in which each sample belongs to the cluster with the nearest mean. Given a set of sample (x1, x2, …, x*n*), where each sample is a *d*-dimensional real vector, then *k*-means clustering aims to partition the *n* samples into *k* sets ($k < n$) S={*S*1, *S*2, …, *Sk*} so as to minimize the within-cluster sum of squares

$$S: \quad avg\ \min\sum_{i-1}^{k} = \Sigma \parallel \mathrm{Xi} - \bar{X} \parallel$$

The most common algorithm uses an iterative refinement technique. The basic step of k-means clustering is simple. K means algorithm will do the three steps below until convergence Iterate until *stable* (= no object move group).

- Determine the centroid coordinate.
- Determine the distance of each object to the centroid.
- Group the object based on minimum distance.

Given an initial set of *k* means $m1^{(1)}$… $mk^{(1)}$, Which may be specified randomly or by some Heuristic, the algorithm proceeds as follows.

Input:

K: the number of clusters.
D: a data set containing n objects.

Output:

A set of clusters.

The algorithm is guaranteed to have converged when the assignments no longer change. Although the K-means method is most widely known and used [8], there are a number of issues related to the method as given below.

- The K-means method needs to compute Euclidean distances and means of the attribute values of objects within a cluster. The classical algorithm is suitable for continuous data.
- The K-means method implicitly assumes spherical probability distributions.
- The results of the K-means method depend strongly on the initial guesses.
- The K-means method can be sensitive to outliers.
- The K-means method does not consider the size of the clusters. Some clusters may be large and some very small.
- The K-means method does not deal with overlapping clusters.

A recent approach to scaling the k-means algorithm is based on the idea of identifying three kinds of regions in data. Regions that are compressible, regions that can be maintained in the main memory, and regions that are discard able. An object is discarding if its membership in a cluster is ascertained. An object is compressible if it is not discarding but belongs to a tight sub cluster. A data structure known as a clustering feature is used to summarize objects that have been discarded or compressed. If an object is neither discard able nor compressible, then it should be retained in main memory [9, 10]. To achieve scalability, the iterative clustering algorithm only includes the clustering features of the compressible objects and the objects that must be retained in main memory, thereby turning a secondary memory based algorithm into a main memory based algorithm [11]. An alternative approach to scaling the k-means algorithm explores the micro clustering idea, which first groups nearby objects into micro clusters and then performs k-means clustering on the micro clusters [12].

IV. **DECISION TREES**

It is flow-chart like tree structure, where each internal node denotes a test on an attribute, each branch denotes an outcome of test, and each leaf node holds a class label. The topmost node in a tree is the root node [14]. Given a tuple, *X*, for which the associated class label is unknown, the attribute values of the tuple are tested against decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision tree is useful because construction of decision tree classifiers does not require any domain knowledge. It can handle hi dimensional data. The learning and classification steps of decision tree induction are simple and fast. Their representation of acquired knowledge in tree form is easy to assimilate by users. Decision tree classifiers have good accuracy [15].

A. Mining Classification Rules

Every data classification project is different but the projects have some common features. Data classification requires some rules [16]. This classification rules are given below: The data must be available, the data must be relevant, adequate, and clean and there must be a well defined problem, and the problem should not be solvable by means of ordinary query while the result must be actionable.

B. Proposed Decision Tree Algorithm

The decision tree algorithm is a top-down induction algorithm. The aim of this algorithm is to build a tree that has leaves that are homogeneous as possible. The major step of this algorithm is to continue to divide leaves that are not homogeneous into leaves that are as homogeneous as possible. Steps of this algorithm are given below.

***Input**:*

Data partition, D, which is a set of training tuples and their associated class labels.

Attribute list, the set of candidate attributes.

Attribute_selection_method, a procedure to determine the splitting criterion that “best” partitions the data tuples into individual classes.

***Output**:*

A decision tree.

C. Decision tree rules

There are a number of advantages in converting a decision tree to rules. Decision tree make it easier to make pruning decisions. Since it is easier to see the context of each rule. Also, converting to rules removes the distinction between attribute tests that occur near the root of the tree and those that occur near the leaves. These rules are easier to read and to understand for people. The basic rules for decision tree are as below.

- Each path from the root to the leaf of the decision tree therefore consists of attribute tests, finally reaching a leaf that describes the class.
- If-then rules may be derived based on the various paths from the root to the leaf nodes.
- Rules can often be combined to produce a smaller set of rules.
- Once all the rules have been generated, it may be possible to simplify the rules.

- Rules with only one antecedent cannot be further simplified. So we only consider those with two or more antecedents.
- Eliminate unnecessary rule antecedents that have no effect on the conclusion reached by the rule.

In some cases, a number of rules that lead to the same class may be combined.

D. Generation of standard decision tree:

For generating decision tree, first we require data table that is given in table-1 as follows:

Table 1: Different attributes

Outlook	Temperature	Humidity	Wind	Play cricket
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Overcast	Mild	High	Strong	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Hot	Normal	Weak	Yes

As, shown in the above table, we see that there are four attributes (e.g. outlook, temperature, humidity, wind) to decide that tennis should be played or not. For result (play cricket), there are two classes such as Yes of No. These attributes may be increased or decreased. But if the numbers of attributes are more than data classification can be done with more accuracy. The decision tree for above data can be generated as shown in figure-1

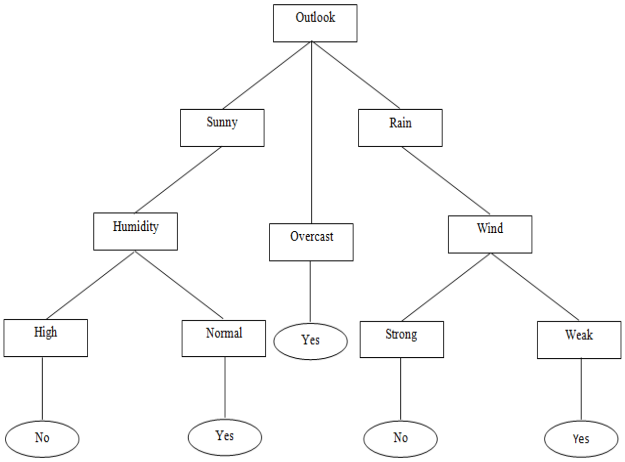

Figure 1: Decision tree for data

V. **CONCLUSION**

Data classification with decision tree is easy when compared with other methods. This research depicts and compares reformulated decision tree with standard decision tree for dataset. The advantage of decision tree is that it provides a theoretical framework for taking into account not only the experimental data to design an optimal classifier, but also a structural behaviour for allowing better generalization capability.

REFERENCES

[1] Hwanjo Yu, Jiong Yang, Jiawei Han, "Classifying Large Data Sets Using SVMs with Hierarchical Clusters", ACM SIGKDD-03, Pp. 24-27, 2003.

[2] H.Zantema and H. L. Bodlaender, “Finding Small Equivalent Decision Trees is hard”, *International Journal of Foundations of Computer Science*; 2000, 11(2):343-354.

[3] Huang Ming, Niu Wenying and Liang Xu,“An improved Decision Tree classification algorithm based on ID3 and the application in score analysis”, *Software Technol.Inst.*,Dalian Jiao Tong Univ., Dalian, China, June 2009.

[4] Chai Rui-min and Wang Miao, “A more efficient classification scheme for ID3”,Sch. of Electron. & Inf. Eng., Liaoning Tech.Univ., Huludao,China; 2010,Version1, pp. 329-345.

[5] IuYuxun and Xie Niuniu “Improved ID3 algorithm”,*Coll. of Inf. Sci.& Eng.*, Henan Univ. of Technol., Zhengzhou,China;2010,pp.;465-573.

[6] Chen Jin, Luo De-lin and Mu Fen-xiang,” An improved ID3 decision tree algorithm”,Sch. of Inf. Sci. & Technol., Xiamen Univ., Xiamen,China, page; 2009, pp. 127-134

[7] Jiawei Han and Micheline Kamber**, “***Data Mining: Concepts and Techniques*”, 2nd edition, Morgan Kaufmann, 2006, ch-3, pp. 102-130.

[8] Safavian, S.R.; Landgrebe, D.; , "A survey of decision tree classifier methodology," Systems, Man and Cybernetics, IEEE Transactions on , Vol. 21, No. 3, Pp.660-674, May/Jun 1991.

[9] Márcio P. Basgalupp, Rodrigo C. Barros, André C. P. L. F. de Carvalho, Alex A. Freitas, Duncan D. Ruiz, "LEGAL-tree: a lexicographic multi-objective genetic algorithm for decision tree induction", SAC '09 Proceedings of the 2009 ACM symposium on Applied Computing.

[10] Carlos Ordonez, "Clustering binary data streams with K-means", DMKD '03 Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery.

[11] Zhexue Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery, Vol. 2, No. 3, 283-304, DOI: 10.1023/A: 1009769707641.

[12] A. K. Jain, M. N. Murty, P. J. Flynn, "Data clustering: a review", ACM Computing Surveys (CSUR) Surveys Homepage archive, Vol. 31, No. 3, 1999.

[13] Watanabe. N, "Fuzzy modeling by hyperbolic fuzzy k-means clustering," Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conferencr, Vol. 2, Pp.1528-1531, 2002 DOI: 10.1109/FUZZ.2002.1006733.

[14] Juanying Xie; Shuai Jiang; , "A Simple and Fast Algorithm for Global K-means Clustering," Education Technology and Computer Science (ETCS), 2010 Second International Workshop on , Vol. 2, Pp. 36-40, March 2010, DOI: 10.1109/ETCS.2010.347.

[15] Du Haizhou; Ma Chong; , "Study on Constructing Generalized Decision Tree by Using DNA Coding Genetic Algorithm," Web Information Systems and Mining, 2009. WISM 2009. International Conference on , Pp.163-167, 7-8 Nov. 2009, DOI: 10.1109/WISM.2009.41

[16] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu and Philip S. Yu, et al., "Top 10 algorithms in data mining", Knowledge and Information Systems, Vol. 14, No. 1, 1-37, DOI: 10.1007/s10115-007-0114-2.

[17] Hang Yang, Fong, S, "Optimized very fast decision tree with balanced classification accuracy and compact tree size," Data Mining and Intelligent Information Technology Applications (ICMiA), 2011 3rd International Conference on, Pp.57-64, 24-26 Oct. 2011.

[18] Guang-Hua Chen; Zheng-Qun Wang; Zhen-Zhou Yu;, "Constructing Decision Tree by Integrating Multiple Information Metrics," Chinese Conference on Pattern Recognition, 2009. CCPR 2009, Pp.1-5, 4-6 Nov. 2009 DOI: 10.1109/CCPR.2009.5344133.