I J C T A, 9(7), 2016, pp. 3359-3365 ISSN: 0974-5572
© International Science Press

A Novel Optimum Depth Decision Tree Method for Accurate Classification

Pullela S. V. V. S. R. Kumar, N. V. Nagamani Satyavani, G. S. N. Murthy* and Jeji Nagendra Kumar Dirisala**

ABSTRACT

The present work derived a classification scheme called Optimum Depth Decision Tree (ODDT) for classification of network audit data collected from KDDCUP 1999 data set. The main problems existed in decision-tree concept used for classification is the presence of a large number of rules and the depth of the tree. The present paper proposed a novel approach ODDT reduces the number of rules and the depth of decision tree considerably. The novelty of the proposed ODDT is that it integrates clustering which is an unsupervised classification with the supervised classification of Decision Trees. The proposed ODDT method is experimented with various clustering methods. The experimental result indicates the efficacy of the proposed method.

Keywords: Decision tree, classification, supervised learning, clustering, unsupervised classification

1. INTRODUCTION

The classification of large data sets is an important problem in area of data mining. The classification approach on a database with a number of records decides the class to which a given record/data belongs to. Many classification algorithms are developed in the literature for the classification of large data sets where supervised and non - supervised classification techniques are two well-known algorithms for an efficient classification of data. Looking into the advantages and disadvantages of both categories of algorithms, the present paper derived Optimum Depth Decision Tree (ODDT) classification scheme. One of the main problems in decision-tree classification is the presence of a large number of rules and the depth of the tree. The proposed novel approach ODDT reduces the number of rules and also the depth of decision tree considerably. The novelty of the proposed ODDT is that it integrates clustering which is an unsupervised classification with the supervised classification of Decision Trees.

The classification can be done in two ways.

1. Supervised classification
2. Unsupervised classification

1.1. Supervised Classification

In supervised classification we follow three major steps i.e. Training phase, Classification phase and Accuracy Assessment phase. In this approach the set of data records available in developing the classification method are divided into two disjoint sub sets i.e. a training data set and a test data set. The training data set is used

* Department of CSE Aditya College of Engineering Surampalem, India, *Email: pullelark@yahoo.com, satyavani363@gmail.com, murthygsnm@yahoo.com*

** Department of IT, Vishnu Institute of Technology Bhimavaram, India, *Email: nagendrakumardj@gmail.com*

in deriving the classifier, while the test data set is used to measure the derived classifier accuracy. The classifier accuracy is determined based on the percentage of test data samples being classified correctly.

One can categorize the attributes of the data into two different types i.e., numerical and non-numerical(categorical) attributes based on attribute domain. There will be a distinguishing attribute which is called as the class label. The goal of the classification is to build a concise model which can predict the class of the records whose label is unknown.

The supervised classification can be implemented in many ways. Several researchers have contributed in this area[5, 13, 14, 15].Three popular methods are discussed in this paper.

1. Classification by Back propagation
2. Bayesian classification
3. Classification by Decision trees

1.1.1. Classification by Back Propagation

Back Propagation is a neural network-learning algorithm. A Neural Network is a set of connected input or output units in which each connection is associated with a weight. In the learning phase, the network learns by adjusting the weights so as to predict the correct class label of the input samples. This approach is also known as connectionist learning.

Though the classification by back propagation has advantages like high tolerance to noisy data, the ability to classify patterns on which they have not been trained it also have some major disadvantages i.e., longer training time, more number of prior input parameters and poor interpretability of symbolic meanings by human beings. These disadvantages made this method less desirable in supervised classification.

1.1.2. Bayesian Classification

Bayesian classifier is a statistical classifier, which can predict class membership probabilities i.e. probability that a given sample belongs to a particular class. Though this classifier has the advantage of minimum error rate, the class conditional independence (effect of an attribute value on a given class is independent of the values of the other attributes) leads to inaccuracies in the classification which leads to limited use of this classifier in supervised classification.

1.1.3. Classification by Decision Trees

The decision tree classifier [1] is a well-known machine-learning techniques and the process of decision tree construction follows divide-and-conquer approach [1]. The decision tree classification scheme generates a tree and a set of rules, representing a model of different classes for the given data set.

Classification of a new record is performed by moving down the tree until a leaf is reached based on the rules. Decision tree classifiers differ in the way it partition the training data into subsets so as to form sub trees as they differ in their criteria for evaluating splits into subsets. The See5 or C4.5 induction algorithm are based on the information theory[2] in evaluating splits. CART method uses Gini Index as a measure for splitting the training samples [3] and some methods use Chi-Square as a measure for splitting.

Based on the studies C4.5 induction decision tree algorithm which based on the Information theory is more accurate and gives reliable results [3, 4, 5] in comparison with other classifiers. The other advantage of C4.5 algorithm is that it can convert decision-tree into corresponding classification rules which are more comprehensive, easy to understand and easy to implement.

1.2. Weaknesses of Decision-Tree Methods

- The problem optimal decision-tree learning [6] is known to be a NP-complete under several aspects of optimality and even for simple concepts. The Practical decision tree learning algorithms which are based on heuristic algorithms such as the greedy algorithm, in which locally optimal decisions are made at each node. These approaches cannot guarantee to return to the globally optimal decision-tree.
- Decision-tree learners create more complex trees which may not generalize the data properly. This is referred as over-fitting and mechanisms such as pruning, which is a complex process, are to be adapted to avoid this problem.

To overcome the above weaknesses, the proposed ODDT method concentrates on reducing the size of decision-tree by integrating it with the unsupervised classification technique like clustering and its representatives without adopting complex pruning techniques.

1.3. Unsupervised Classification

In unsupervised classification class label of each training data sample is not known and also the number of classes to be learned may not be known in prior. Clustering is one of the highly used techniques in unsupervised classification. There are many clustering techniques in the literature including hierarchical clustering [10, 11], self-organizing maps [12] and partitioned algorithms and they have been used for various purposes. In many applications, the purpose of clustering is to find a concise representation for a large set of examples by analyzing and determining a structure among the data and fitting a set of clusters to them. If one can gather all similar examples into one group and find some way to represent the group as a whole, then one can summarize the information for all the data examples in that cluster. A more precise and fine-grained representation of the overall dataset can be achieved by a greater number of clusters. One can enjoy the greater savings in the compression of data representation, at the cost of losing finer details about the dataset by fewer clusters. However, different types of cluster representations may be used to summarize the cluster.

Cluster representatives, appropriately enough, represent the clusters they are selected for. They are the means by which one may summarize the examples within the clusters. Therefore, it is important to not only have clusters that are indeed collections of examples (intervals) with strong similarities (behavior), but also to choose cluster representatives that summarize all the examples in the cluster well.

2. A NOVEL APPROACH TO CONSTRUCT OPTIMUM DEPTH DECISION TREE (ODDT) BASED ON CLUSTER REPRESENTATIVES

First, In Iterative Decotomizer 3 (ID3) decision-tree construction algorithm discretization is a prerequisite step in case of continuous values. This discretization process can be done using binning in ID3. The main problem in this method is selection of bin size which is data- independent. The binning procedure will not consider inherent property of data to form bins. The discretization using clustering uses the inherent property of data and will reach the goal of reducing the size of decision-tree. In case of C4.5 this discretization step is not required in continuous values. But implementing clustering and then implementation of C4.5 using cluster representatives to construct decision-tree will reach the goal of reducing the number of rules and depth of the decision tree.

The proposed ODDT algorithm 1 uses three clustering methods, namely, basic K-means clustering algorithm and two of recent algorithms called as Perimeter K-Means (PKM) clustering algorithm [4, 6] which picks two data points at a time and Weighted Interior K-means Clustering (WIKC) algorithm[7] in which gridding and construction of mountain function is completely eliminated and clusters are formed automatically without prior knowledge of K value and other parameters.

Algorithm 1: To Construct Optimum Depth Decision Tree(ODDT).

```
Begin
Step 1  :  Select training set from data set.
Step 2  :  Cluster using clustering algorithm.
Step 3:    Calculate cluster representatives from each cluster.
Step 4  :  Construct Decision Tree using C4.5 Decision Tree algorithm from cluster representatives.
Step 5  :  Verify the accuracy of classifier using test data set.
End
```

The cluster representative plays a vital role in increasing the efficiency of the classifier. For this the proposed ODDT approach uses five Cluster Representative Types (CRT) to construct an efficient and precise Decision Tree.

CRT-1 is the random point from frequently occurring class. The proposed ODDT assumes that similar data points occur frequently in the cluster, due to their frequency of occurrence in the cluster group. The novelty of the proposed scheme is that from the above frequently occurring data points, the proposed ODDT randomly selects one data- point as a cluster representative. In CRT-2, representatives are calculated as in CRT-1 but in every class random representatives are considered from the clustered data points. CRT-3 is a point which is nearer to the center using distance metrics. CRT-4 is a point which is closest to the center of a cluster, CRT-5 is the point that is closest to the overall, or average behavior of the examples in that cluster. The point x with the minimum distance from the cluster centers *c* is chosen as the representative point.

The proposed ODDT method examines the Euclidean distance between data-point x and centroid *c* given in 1.

$$d(\mathrm{x}, \mathrm{c}) = |\mathrm{x}\text{-}\ c\ | = \sqrt{\sum_{i=1}^{n}(x_i - c_i)^2} \tag{1}$$

Where x is a point in the cluster and c, centroid or cluster center calculated using mean and n number of attributes or dimension of data set. CRT-4 is the point which is nearer to center from frequently occurring Class. The CRT-4 is based on another novel concept, namely, choosing the cluster representative points which are nearer to the center as in CRT-3 from the frequently occurring data-points. CRT-5 is a point which is nearer to the center from every Class. The novelty in this CRT-5 is that the nearest points are considered as in CRT-3 but they are from every class in the cluster.

3. EXPERIMENTAL ANALYSIS

The proposed method is applied to network audit data collected from KDDCUP 1999 data set [8]. The number of such records is 172 with 6 features which are service, protocol, flag, source bytes, destination bytes and duration. To implement the proposed ODDT algorithm this data-set of 172 records has to divide into training data set and test data set. The ratio of training to testing set will differ from application to application. For our experimentation, these 172 records were divided into 66.86% as training set and 33.13% as test set. Hence 115 records as training set and 57 records as test set. These were selected randomly from the 172 records.

The proposed ODDT includes three clustering techniques which are WIKC, PKM and K-means. It is tested with the highest K values which are 6 in K-means, 9 in PKM and 18 in WIKC. Five types of cluster

Table 1: Comparison of C4.5 And Oddt

		Optimum Depth Decision Tree(ODDT)														
		K-Means					*PKM*					*WIKC*				
	C 4.5	*CRT-1*	*CRT-2*	*CRT-3*	*CRT-4*	*CRT-5*	*CRT-1*	*CRT-2*	*CRT-3*	*CRT-4*	*CRT-5*	*CRT-1*	*CRT-2*	*CRT-3*	*CRT-4*	*CRT-5*
Number of Clusters		6	6	6	6	6	9	9	9	9	9	18	18	18	18	18
Number of cluster repres-entatives as training set	No. of training records 115	6	15	6	6	15	9	24	9	9	24	18	24	18	18	24
Number of rules generated by Decision Tree	282	27	20	18	18	13	25	17	16	15	12	24	16	14	13	10
Number of Test records	57	57	57	57	57	57	57	57	57	57	57	57	57	57	57	57
Number of Test records correctly classified	48	48	48	50	51	51	48	49	50	51	52	50	51	51	52	53
Accuracy	84.21	84.2	84.2	87.71	89.47	89.47	84.21	85.96	87.71	89.47	91.22	87.71	89.47	89.47	91.22	92.28
Depth of the Decision Tree	7	5	5	4	3	3	5	4	4	3	3	4	3	3	3	2

representatives are denoted by CRT-1 to CRT-5 are considered to represent clusters formed by WIKC[7], PKM[9] and K-means algorithms. The ODDT algorithm constructs the decision-tree with representatives of clustered training data set and tested with test data set. This proposed ODDT is compared with C4.5 and the results are tabulated in Table 1.

From Table 1 it is evident that in the classification accuracy by C4.5 is 84.21% which is less than the proposed ODDT method with 91.22%. In the proposed ODDT method the tests are made with three clustering techniques PKM, K-means and WIKC to form clusters and its representatives. The results show that the accuracy depends on the number of clusters and its representative types. The K-means can form a maximum of 6 clusters from the data set and its accuracy for some representative types is low when compared to PKM which has a maximum of 9 clusters, whereas WIKC forms 18 clusters giving the highest classification accuracy.

From Fig.1, it is evident that in the proposed ODDT, the highest accuracy is given by CRT-5 representatives which are nearest to the center in every class and CRT-4 representatives which are nearest to the center in frequently occurring classes giving the next highest accuracy. The third best performance is given by CRT-3 cluster representatives which are single points in every cluster and are very near to cluster center. The CRT-1 representatives are randomly selected from frequently occurring classes and CRT-2 representatives are also selected randomly from those which frequently occurred in every class, showing poorer performance than the other types. Hence it is proved that random selection may not give the correct result in most of the cases. Hence the accuracy of the classifier with the representatives which are very near to the center in every class: this is CRT-5 giving the highest accuracy and optimum depth-decision tree. The results are also shown in graphs of Fig. 1 to 3.

The above tables and graphs clearly indicate the fact that the proposed ODDT reduces the number of rules and depth of decision- tree compared to C4.5.

4. CONCLUSION

The present paper studies various supervised and unsupervised classification techniques but concentrates on the supervised classification. One of the main problems in decision-tree classification is the presence of

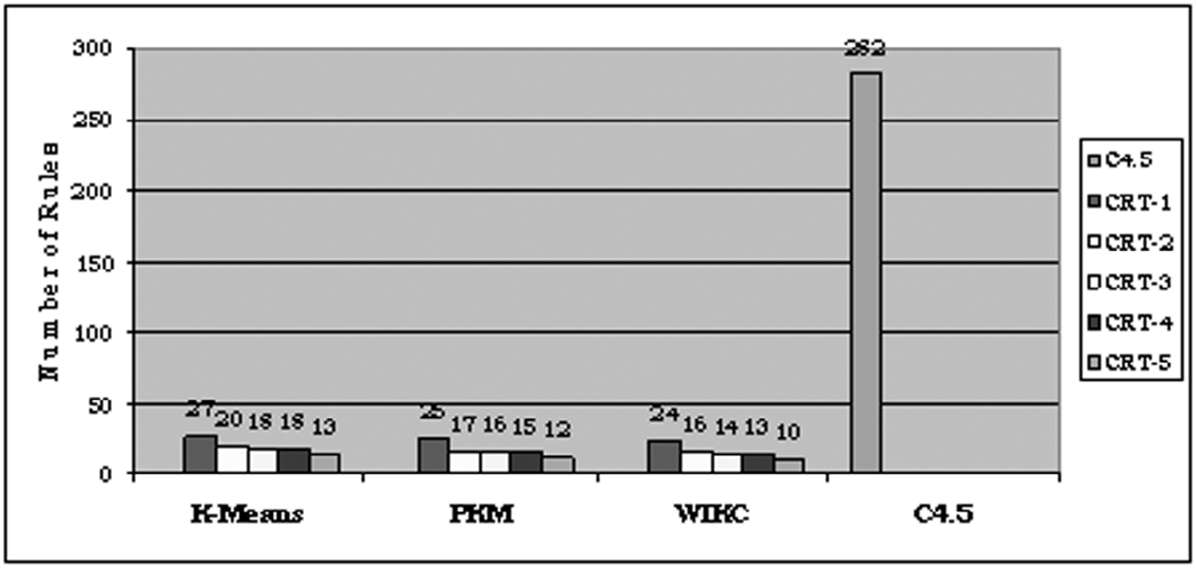

Figure 1: Comparison of Number of Rules.

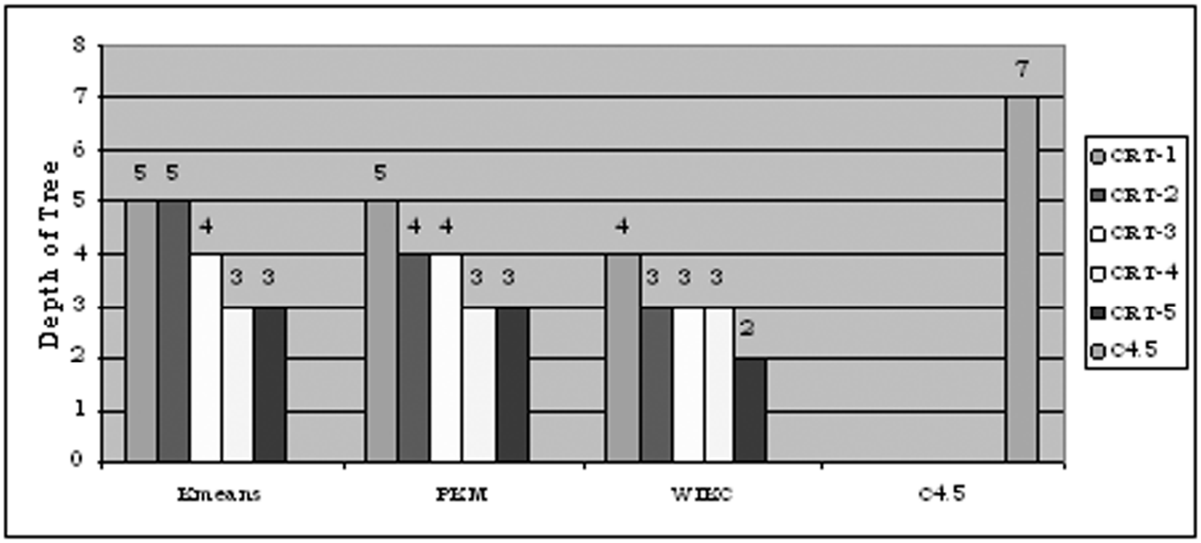

Figure 2: Comparison of Depth of Tree

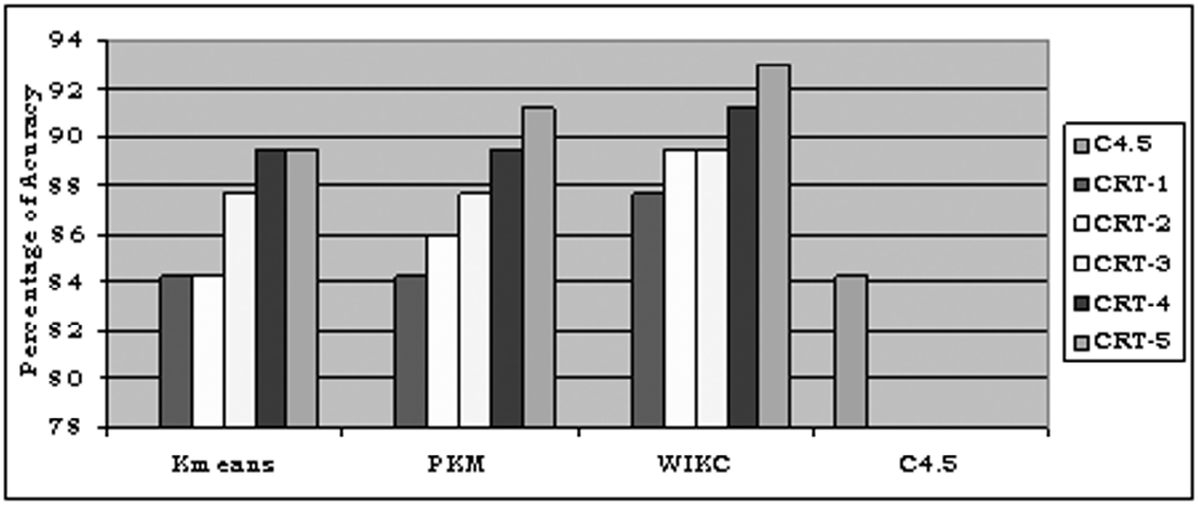

Figure 3: Comparison of Percentage of Accuracy

a large number of rules and the depth of the tree. The proposed novel approach ODDT reduces the number of rules and the depth of decision tree considerably. The proposed decision-tree classifier is trained using the training set and tested with the test set from KDDCUP99 Intrusion Detection (ID) data set. The novelty of the proposed ODDT is that it integrates clustering which is an unsupervised classification with the supervised classification of Decision Trees. This proposed ODDT is tested with the test data and found that the accuracy of this classifier is more than the accuracy of direct implementation of C4.5 classifier and also in terms of time consumption. This leads to our goal of reducing the number of rules and the depth of tree without applying any complex pruning techniques. The resultant classifier trained by our proposed ODDT classifies the unknown record with more accuracy and consuming less time.

REFERENCES

[1] Quinlan. J.R., "Programs for Machine Learning," California: M.K. Publishers Inc.tp://www.cs.bris.ac.uk/~peng/publications/SoftDT.pdf New York: John Wiely and Sons, 1993.

[2] Shannon C. and W. Weaver W., "The Mathematical Theory of Communication," USA: University of Illinois Press, 1949.

[3] Agrawal.R., Gehrke. J., et al. "Automatic subspace clustering for high dimensional data for data mining applications," SIGMOD-98, 1998.

[4] Eklund P.W., Kirkby S.D. & A. Salim, Data mining and soil salinity analysis," International Journal of Geographical Information Science, Vol. 12, pp. 247-268, 1998.

[5] German.G.W.H., West G. and Gahegan M., Statistical and AI Techniques in GIS Classification, " A Comparison, site: http://divcom.otago.ac.nz/sirc webpages/99German.pdf (accessed on 12-05-2004), 1999.

[6] Lin C.R., and M.S. Chen, "On the Optimal Clustering of Sequential Data", In Proceedings of Second International Conference on Data Mining, April 2002.

[7] GVSNRV Prasad et al., "Automatic Clustering based on initial seed", vol. 3 no. 12 Dec. 2011 pno 3800-3806.

[8] Fayyad U., "Data Mining and Knowledge Discovery in Databases: Implications from scientific databases", In Proceedings of 9th International Conference on Scientific and Statistical Database Management, pp. 2-11, Olympia, Washington, USA, 1997.

[9] GVSNRV Prasad, Ch. Satyanarayana, Vijay kumar V., "Perimeter clustering algorithm to reduce the no of iteration," IJCA Vol. 3 no. 8 Dec 2011 pp. 41-46

[10] Anderberg, M.R. "Cluster Analysis for Applications," Academic Press, New York, 1974.

[11] Ward. J.H., "Hierarchical Grouping to Optimize an Objective Function," Journal of the American Statistical Association, Vol. 58., No. 301. pp: 235–244, 1963.

[12] Boujemaa .N, "On Competitive Unsupervised Clustering," International Conference on Pattern Recognition (ICPR'00) Vol. 1, pp. 1631-1634, 2000.

[13] D. J. Nagendra Kumar, J. V. R. Murthy, Suresh Chandra Satapathy and S.V.V.S.R. Kumar Pullela. A Wrapper Approach of Feature Selection Using GP Ensemble Classifier. International Conference on Swarm, Evolutionary and Memetic Computing (SEMCCO- 2010).

[14] D. J. Nagendra Kumar, J.V.R. Murthy, Suresh Chandra Satapathy, and S.V.V.S.R. Kumar Pullela, "A Study of Decision Tree Induction for Data Stream Mining Using Boosting Genetic Programming Classifier", B.K. Panigrahi et al. (Eds.): SEMCCO 2011, Part I, Lecture Notes in Computer Science, Springer Berlin Heidelberg, Volume 7076, pp 315-322, 2011.

[15] D. J. Nagendra Kumar, Suresh Chandra Satapathy and J. V. R. Murthy. A Scalable Genetic Programming Multi-class Ensemble Classifier. IEEE World Congress on Nature & Biologically Inspired Computing (NaBIC-2009), pp. 1201-1206.