International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Accuracy Improvement of C4.5 using K means Clustering

Driyani Rajeshinigo[1], J. Patricia Annie Jebamalar[2]

[1]Research Scholar, Dept. of Computer Science, St.Xavier's College, Tirunelveli, Tamilnadu, India

[2]Assistant Professor, Dept. of Computer Science, St.Xavier's College, Tirunelveli, Tamilnadu, India

Abstarct: ***Data mining is used to extract useful information from the vast amount of data. Classification is one of the techniques of data mining which will be used to predict the target attribute accurately from the knowledge it gained from the training data. There are various classifiers which among decision tree is very simple and most effective method. Accuracy of a classifier is how well it predicts the target attribute of test data correctly from the knowledge gained from the training set. In this paper, classification accuracy of C4.5 is improved with K means Clustering and adding the tested data dynamically to the training set. This improved C4.5 predicts the target variable with higher accuracy level with the help of K means clustering which is used to discretize the continuous data.***

Keywords: Decision tree, C4.5, K-Means, Continuous data

1. Introduction

Data mining is used to analyze large amounts of data effectively to discover some useful information. Classification is one of the techniques of data mining which will be used to predict the target attribute accurately from the knowledge it gained from the training data.

There are various classifiers such as Decision trees, Bayesian classifier, random forest, neural network and support vector machine. Among all, decision tree is widely used because it is very simple, easy to understand and most effective method.
Clustering is an unsupervised classification which is used to group the similar values without target attribute. There are various clustering approaches such as hierarchical, partitioning, grid based, density based and model based. Among all, partitioning based K means clustering method is very simple. Hybrid of C4.5 with K means is used in various applications like networking, medicine and educational mining.

In this paper, C4.5 accuracy is improved by integrating with k means clustering where this clustering is used to handle the continuous data. This continuous data brings disadvantages to the decision tree which ultimately lowers the accuracy.

2. Related Work

This section summarises literature review of various studies made on C4.5, K means clustering and the integration of C4.5 with K means clustering.

C4.5 is optimized using L'hospital rule instead of logarithmic function which takes longer time when calculating gain ratio. L'hospital rule works faster and gives more effective results than the normal one [8]. The data mining tool WEKA has been used as an API of MATLAB for generating the modified C4.5 classifier. C4.5 has been improved for Diabetes data set by Loading ARFF file from WEKA to MATLAB and refine the data set using MATLAB then C4.5 is applied which improved the accuracy significantly [2]. Decision tree algorithm and cluster analysis is integrated to classify a given data set. Decision tree or clustering is considered for classification based on the information gain value. The algorithm whichever has the greater information gain is being selected for the given data set [5]. Classification algorithms are studied on a student set and found C4.5 is highly used for predicting student's academic performance [1]. Simple operations of algebra are used in C4.5 instead of log function which takes long time to calculate information gain. This improvement of using algebraic operations consume less time than the existing one [9]. C4.5 is improved by applying the feature selection to reduce the dataset dimension and reduced error Pruning to remove the sections of the tree that provide little power to Classify instances and then applied cross validation methods which overall improved classification accuracy [10]. Decision tree algorithm C4.5 is improved by applying post running to remove the branches of the tree which gives least effect for classifying the instances and then proper cross validation methods were applied which improved the accuracy of the algorithm [11].

K means Clustering algorithm calculate the distance between each data object and all cluster centers in each iteration, which makes the efficiency of clustering is not high. This paper proposes an improved k-means algorithm which requires a simple data structure to store some information in every iteration, which is to be used in the next interation. The improved method avoids computing the distance of each data object to the cluster centers repeatedly which effectively improve the speed of clustering and accuracy, reducing the computational complexity of the k-means [14]. K means clustering is very sensitive to the initial values. This Paper proposes new algorithm to improve this scenario based on Iterative density. Through continuous modification to density threshold, it gets the more clustering centers and merges them until the specified number of clustering center is met which optimizes the dependence[7]. This paper optimizing the running time of K means clustering and it comes from the observation that after a certain number of iterations, only a small part of the data

Licensed Under Creative Commons Attribution CC BY

elements change their cluster, so there is no need to re-distribute all data elements. Therefore the paper puts an edge between those data elements which won't change their cluster during the next iteration and those who might change it, reducing significantly the workload in case of very big data sets [3].K means clustering is used for the defect segmentation of fruits. The pixels are clustered based on their color and spatial features then the clustered blocks are merged to a specific number of regions which provides a feasible solution for defect segmentation [6].

Hybrid of K means clustering with C4.5 is used for classifying the anomalies and normal activities in the network system. The K means clustering algorithm is used to partition the training instances into K clusters and build C4.5 decision tree for each cluster which refines the decision boundaries by learning the subgroups from the cluster [13]. Node Localization in a wireless network has been handled by k means clustering with the decision tree C4.5.The combination of both algorithms has been used to choose the best node among a set of intersection points from the anchor and localized node within the range of an unlocalized node [12]. Hybrid of K means clustering with C4.5 is used for classifying the breast cancer. Malignant cases are extracted from the dataset and given to the k means clustering to partition the instances into clusters and Decision tree is build for each cluster. This hybrid algorithm provides the results which will be very useful to experts to diagnose malignant cases with better accuracy and reliability to find the type of cancer [4]. K means clustering and decision tree is used to promote the customer value. This investigation first applies the K means method to divide the customers into high, middle and low valued groups. Decision tree is utilized to mine the characteristics of each customer segment which will be the valuable reference for the managers to promote customer value [15].

3. Methodology

3.1. Clustering:

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters .It is an unsupervised classification.

3.1.1. K means Clustering:

K-means is a partition based clustering method that aims to find the positions of the clusters that minimize the distance from the data points to the cluster. **K** is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. The numbers of clusters are fixed at first. The main idea is to define K centroids for each cluster. The next step is to associate each point in the data set to the nearest centroid. When no points are in data set, the grouping is done. Now New centroids have to be found and a new building has to be done between the same data point and the nearest new centroid. As a result K centroids change their location step by step until no more changes are done.

Algorithm: K means Clustering
1. Select K points as the initial centroids
2. **Repeat**
3. Form K clusters by assigning all points to the closest centroid
4. Recompute the centroid of each cluster
5. **until** the centroids don't change

3.2. Classification

Classification is one of the Data Mining techniques that are mainly used to analyse a given dataset. It is used to extract models that accurately define important data classes within the given dataset.
Classification is a two step process.
Step 1: The model is created by applying classification algorithm on training data set
Step 2: The extracted model is tested against a predefined test dataset to measure the model trained performance and accuracy.
So classification is the process to assign class label from dataset whose class label is unknown.

3.2.1. Decision Tree

A decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labelled with distinct outcomes of the test. Each leaf node has a class label associated with it.

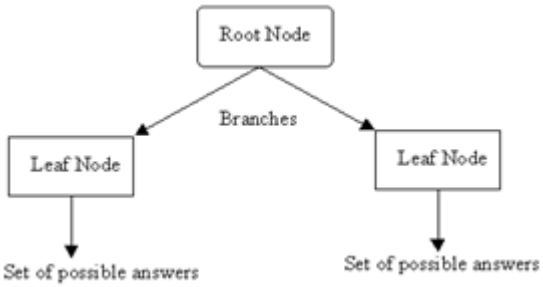

Figure 1: Decision Tree

A. C4.5:
This algorithm is a successor to ID3 developed by Quinlan Ross .C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. C4.5 uses Entropy and Information gain as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute.
Entropy is the measure of disorder or impurity

$$\text{Entropy} = -\sum_{i} P_i \log_2 P_i$$

P_i is the probability of class i

Information gain tells us how important given attribute of the feature vector is. This information gain is used to decide

Licensed Under Creative Commons Attribution CC BY

the ordering of attributes in the nodes of a decision tree. More precisely the information gain, Gain(S, A) of an attribute A, relative collection of examples S, is given by equation.

$$\text{Gain}(A) = I(S1,S2, \cdots, Sm) - E(A)$$

In other words gain (A) is the expected reduction in entropy caused by knowing the Value of attribute A.

At first, calculate Entropy and information gain for each attribute. The root node will be the attribute whose information gain is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

Algorithm C4.5:

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

Pseudo code:

In pseudo code, the general algorithm for building decision trees is

1) Check for the above base cases.
2) For each attribute *a*, find the normalized information gain ratio from splitting on *a*.
3) Let *a_best* be the attribute with the highest normalized information gain.
4) Create a decision *node* that splits on *a_best*.
5) Recur on the sublists obtained by splitting on *a_best*, and add those nodes as children of *node*.

4. Accuracy Improvement of C4.5

C4.5 is the most popular classification algorithm. There are many advantages of using C4.5 such as accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. The drawback of this classifier is that it gives lower accuracy with continuous data. Accuracy of a classifier is how well it predicts the target attribute of test data correctly from the knowledge gained from the training set. In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it. Another problem is that continuous attribute are being considered more than once while building the tree which results in a larger tree leads to post pruning.

4.1Transforming continuous attribute

C4.5 algorithm reads the training data and builds the decision tree. While constructing the tree the algorithm checks whether any continuous attributes are given in the data set. If yes, the algorithm calls the K means clustering algorithm with the attribute values. K means clustering is used here to transform the continuous values of the attribute as categorical value. Once the values are categorized, k means clustering returns the categorical value for that particular attribute to C4.5 algorithm. Then C4.5 considers these categorical values for that attribute and builds the tree. This process of transforming continuous values into categorical values using K means clustering while constructing the tree greatly improves the accuracy of the classifier.

5. Experimental Results

The Proposed work has been evaluated with the students' data set. Educational mining is the current trend which uses data mining. There is huge amount of data stored in educational database about students but being unused. Educational institutions normally execute some queries on database to fetch past records about a student. But the data stored in educational database can predict a student's performance if used correctly. This can help the student to improve himself in future and can help the staffs to give some additional care for the students who were not performing well enough. Choosing the attributes from the data set for classification lays vital role to predict the target attribute accurately. The students data set used to evaluate the performance of the proposed work is given below

Table 1: Students Variables

Variable	*Description*	*Possible Values*
IAT	Internal Assessment Test	Numeric
AS	Attendance Status	{High, Low}
CA	Current Arrears	{Yes, No}
HTL	Hostel	{Yes, No}
PC	Project Completion	{Yes, No}
AC	Assignment Completion	{Yes, No}
PW	Parents Work	{Yes, No}
AD	Academic Detention	{Yes, No}

This data set has been given to the C4.5 algorithm which starts building the Decision tree. When it takes the attribute IAT, C4.5 calls K means clustering which transforms this continuous data into categorical data. K means clustering initiates the number of clusters to 2, one cluster is "satisfactory" and another one is "not satisfactory". The IAT average which is greater than or equal to 50 falls on the category "satisfactory" and the values which is less than 50 falls on the category "Not satisfactory". K means clustering returns the categorical value to the decision tree and then the algorithms C4.5 resumes its work to build the tree. This hybrid of C4.5 with K means cluttering increase the accuracy from around 73% to 92%.

The below Table 2 will show the accuracy obtained with the existing C4.5 algorithm and the Proposed one

Table 2: Classifier Accuracy

Algorithm	Accuracy
C4.5	73%
K means clustering with C4.5	92%

The below Figure 1 column chart shows the graphical representation of Table 2

Licensed Under Creative Commons Attribution CC BY

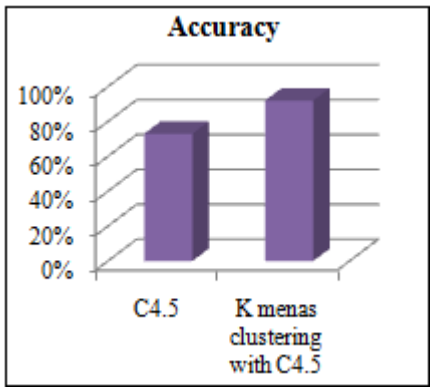

Figure 1: Classifier accuracy

6. Conclusions and Future Work

C4.5 is the most popular classification algorithm. This algorithm has the disadvantage of handling continuous attributes which greatly reduces the classifier accuracy. Transforming the continuous values of the attribute into categorical value using K means clustering while building the decision tree improves the classification accuracy significantly. In addition to that, the tested data using the proposed work will be added to the training set which again improves the accuracy of the classifier. As a future work, The Proposed work can be tested with other data sets and other clustering algorithms can be used instead of k means clustering.

References

[1] Amirah mohamed Shahiri, Wahidah Husain, "A Review on Predicting Student's Performance Using Data Mining Techniques", ELSEVIER, Volume 72 , Pages 414-422, 2015.
[2] Gaganjot Kaur and Amit chhabra, "Improved C4.5 Classification Algorithm for the prediction of Diabetes", International Journal of Computer Applications, Volume 98 – No.22, July 2014.
[3] Cosmin Marian Poteras, and Mihai Mocanu "An Optimized Version of the K-Means Clustering Algorithm", Conference on Computer Science and Information Systems, 2014
[4] Walid Meliani and Nahla Ben Amor, "A Hybrid Approach Based on Decision Trees and Clustering for Breast Cancer Classification", International Conference of Soft Computing and Pattern Recognition, 6th International Conference of IEEE, 2014
[5] Masaki KUREMATSU and Hamido FUJITA, "A Framework for Integrating a Decision Tree Learning Algorithm and Cluster Analysis", 12th IEEE International Conference on Intelligent Software Methodologies, Tools and Techniques, September 22-24, 2013
[6] Shiv Ram Dubey , Pushkar Dixit and Jay Prakash Gupta, "Infected Fruit Part Detection using K-Means Clustering Segmentation Technique", International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 2, 2013.
[7] Liu Guoli, YuLimei and Gao Jinqiao. "The Improved Research on K-Means Clustering Algorithm in Initial Values", International Conference on Mechatronic Sciences, Electric Engineering and Computer, December 2013.
[8] Gaurav L. Agrawal and Hitesh Gupta, "Optimization of C4.5 Decision Tree Algorithm for Data Mining Application", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 3, March 2013
[9] Ping Gu and Qi Zhou, "Student Performances Prediction Based on Improved C4.5 Decision Tree Algorithm", Emerging Computation and Information teChnologies for Education. Springer Berlin Heidelberg, 2012.
[10] A. S. Galathiya, A. P. Ganatra and C. K. Bhensdadia, "Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning", International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012
[11] A. S. Galathiya, A. P. Ganatra and C. K. Bhensdadia, "Classification with an improved Decision Tree Algorithm", International Journal of Computer Applications, Volume 46– No.23, May 2012
[12] Khalid K. Almuzaini and and T. Aaron Gulliver, "Localization in Wireless Networks using Decision Trees and K-means Clustering", Vehicular Technology Conference (VTC Fall), 2012 IEEE. IEEE, 2012.
[13] Amuthan Prabakar Muniyandi and R. Rajeswari, "Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm", International Conference on Communication Technology and System Design, 2011
[14] Shi Na and Liu Xumin, "Research on k-means Clustering Algorithm", Third International Symposium on Intelligent Information Technology and Security Informatics, 2010
[15] Yi-Hui Liang, "Combining the K-means and decision tree methods to Promote Customer Value for the Automotive Maintenance Industry", Industrial Engineering and Engineering Management, 2009

Author Profile

Driyani Rajeshinigo is a Research Scholar in the Computer science Department, St.Xavier's College, Tirunelveli. She received Master of Computer Science (M.Sc) degree in 2007 from St.Joseph's College, Trichy. Her research interests are from Data Mining.

J. Patricia Annie Jebamalar is an Assistant professor in Department of Computer Science, St.Xavier's College, Tirunelveli. She received her Master of philosophy (M.Phil) in Computer science from Alagappa University, Karaikudi. She has published more than 9 Research Papers in Journals and Conferences. Her research interests are from Data Mining.

Licensed Under Creative Commons Attribution CC BY