TRANSFORMING TECHNOLOGY
EXPLORING KNOWLEDGE

Volume 2 Issue 3
July 2014 - September 2014
ISSN: 2322-0821(online)
ISSN: Applied for (Printed)

INTERNATIONAL
RESEARCH JOURNAL
OF ENGINEERING
& APPLIED SCIENCES
www.irjeas.com

C-Means Unsupervised Classification Using CART

Shazia Sultan[1], Amar Nayak[2]

[1]*M.Tech Scholar, Computer Technology & Application, TIT, Bhopal, INDIA 462021*
[2]*Head, Department of Computer Science Engineering, TIT, Bhopal, INDIA 462021*

Abstract **- Data Mining is a field of search and researches of data. Mining the data means fetching out a piece of data from a huge data block. The basic work in the data mining can be categorized in two sub sequent ways. One is called classification and the other is called clustering. Although both refers to some kind of same region but still there are differences in both the terms. The classification of the data is only possible if you have modified and identified the clusters. In the field of data mining, classification is the very important techniques to find out new patterns. K-nearest neighbor algorithm is most usable methods of classification. However, individually they face few challenges, such as, time utilization and inefficiency for very large databases. The current paper attempts to use the K-nearest neighbor algorithm and Clustering methods for classification of data mining. We have made some changes in K-nearest neighbor algorithm and used it to classify data. This technique helps in improving the efficiency of KNN to a high extent.**

I. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amounts of data. Classification (technique to analyses the *frequent item sets*) is one of the major fields in the area of extracting knowledge from vast data. A *frequent item set* typically refers to a set of items that frequently appear together in a transactional data set, such as milk and bread.

Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision-making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the "essence" of information stored, and the discovery of patterns in raw data.

The general field of data mining encompasses a range of tasks, each specific to a different set of target problems. Major areas of data mining are as follows:

- **Summarization:** the process of summarization is that of finding a simple model or description for a given set of data. Some examples of this might be finding summary rules to describe common properties for a subset of data, or visualization methods for displaying relationships within the data.

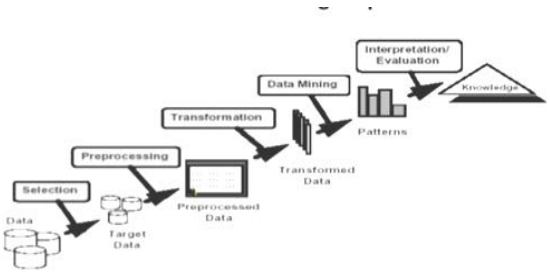


Fig. 1.1 – Data Mining Process

- **Deviation detection:** this process is used to detect the significant and consistent changes in a set of data over a series of time steps. The process derives a model which defines which attributes change and how they change over time.

- **Classification:** this is the task of finding some model or function which maps each element of data to one class, out of a discrete set of possible classes. Each element of data must belong to only one class, and the classification task is to accurately assign each element of data to the appropriate class.

- **Regression/ Numeric Prediction:** similar to classification the task of numeric prediction is to derive some model or function which maps each element of data to a value. In this case however

International Research Journal of Engineering & Applied Sciences
www.irjeas.com, ISSN: 2322-0821, Volume 2, Issue 3, July 2014 – September 2014, Page 67-71

the prediction value is a continuous value. A model must predict an output value to the appropriate value as possible for each element of data.

- **Association Finding/Dependency Modeling:** this problem involves the derivation of a model which describes dependencies or associations between variables of a problem. The model will indicate that on variable, or a set of variables, influences the value of another variable, and will often indicate the strength of this influence.

- **Clustering:** the task of clustering is to find some model which distinguishes a subset of data elements from the main body of data elements based upon some characteristics of the subset. The elements which form the cluster should be closely related in some way, in which data elements outside the cluster are not related. The measure of how closely related elements of the cluster are and how greatly this differs from the wider body of data elements indicates the strength of the cluster.

Fig. 1.2 – Data Mining Cycle

II. WORK ALREADY DONE

Here we will briefly describe the previously work already done on the algorithm which is used in dissertation as following

C4.5 ALGORITHM

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms.

The C4.5 algorithm is slightly work differently, C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S = s1,s2,... of already classified samples. Each sample si= x1,x2,... is a vector where x1,x2,... represent attributes or features of the sample. The training data is augmented with a vector C = c1,c2,... where c1,c2,... represent the class to which each sample belongs. C4.5 belongs to a succession of decision tree learners that trace their origins back to the work of Hunt and others in the late 1950s and early 1960s (Hunt 1962)[42]. Its immediate predecessors were ID3 , a simple system consisting initially of about 600 lines of Pascal, and C4 . C4.5 has grown to about 9,000 lines of C that is available on diskette.

Input to C4.5 consists of a collection of training cases, each having a tuple of values for a fixed set of attributes (or independent variables) A = fA1; A2; :::; A kg and a class attribute (or dependent variable). An attribute Aa is described as continuous or discrete according to whether its values are numeric or nominal. The class attribute C is discrete and has values C1; C2:Cx. The goal is to learn from the training cases a function DOM (A1) fi DOM (A2) fi:fi DOM (Ak)! DOM(C) that maps from the attribute values to a predicted class. The distinguishing characteristic of learning systems is the form in which this function is expressed. A decision tree is depicted as a recursive structure of a leaf node labeled with a class value, oral test node that has two or more outcomes, each linked to a sub tree. The work that already have performed using this technique is referenced by [2][3][4].

CART ALGORITHM

Classification and Regression Trees is a classification method which uses historical data to construct so-called decision trees. Decision trees are then used to classify new data. In order to use CART we need to know number of classes a priori. CART methodology was developed in 80s by Breiman, Freidman, Olshen, and Stone. For building decision trees, CART uses so-called learning sample - a set of historical data with pre-assigned classes for all observations. For example, learning sample for credit scoring system would be fundamental information about previous borrows (variables) matched with actual payoff results (classes).Decision trees are represented by a set of questions which splits the learning sample into smaller and smaller parts. CART asks only yes/no

International Research Journal of Engineering & Applied Sciences
www.irjeas.com, ISSN: 2322-0821, Volume 2, Issue 3, July 2014 – September 2014, Page 67-71

questions. A possible question could be: "Is age greater than 50?" or "Is sex male?" CART algorithm will search for all possible variables and all possible values in order to find the best split – the question that splits the data into two parts with maximum homogeneity. The processes then repeated for each of the resulting data fragments. Here is an example of simple classification tree, used by San Diego Medical Centre for classification of their patients to different levels of risk:

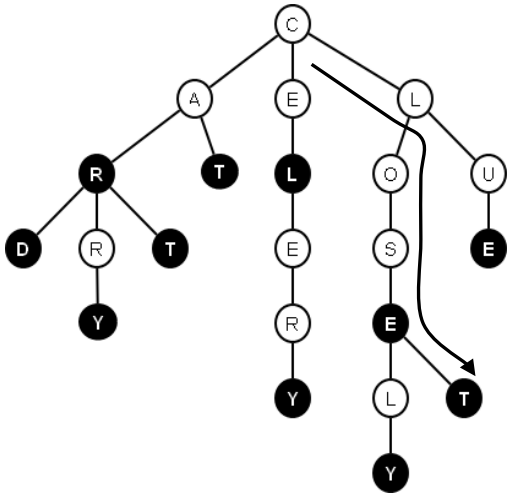


Fig. 2.1 – Tree Construction

III. WORKS TO BE CARRIED OUT

1. Study the C4.5 algorithm and CART algorithm which is used for classifying the dataset.
2. Study the C-mean algorithm which is used for making the cluster of concerned dataset.
3. Find an easy and perfect algorithm which can implement better result than the proposed algorithms till now.
   For this we will uses result of two different algorithms CART and C4.5.
4. Implement the algorithm in order to classify the dataset after cluster region is being defined using C-mean technique.
5. We can improve the accuracy of classifying data using the feature of both techniques compared to the previous available technique of classifying the data.

IV. DATA MINING AND KNOWLEDGE DISCOVERY

Data mining and knowledge discovery in database have been attracting a significant amount of research, industry and media attention of late. What is all the excitement about? This section provide an overview of this emerging field, classifying how data mining and knowledge discovery in database are related both each other and to related field, such as machine learning, statics and databases. This section mentioned particular real world application, specific data mining technique. Challenges involved in real world application of knowledge discovery and current and future research direction in the field.

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge.

- **Data cleaning:** also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

- **Data integration:** at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

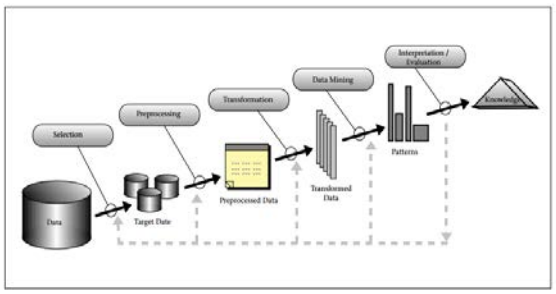


Fig. 4.1 – Steps of discovering knowledge from raw data

- **Data selection:** at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

- **Data transformation:** also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure]

International Research Journal of Engineering & Applied Sciences
www.irjeas.com, ISSN: 2322-0821, Volume 2, Issue 3, July 2014 – September 2014, Page 67-71

V. CONVENTIONAL ALGORITHMS OF DATA MINING

There are some conventional algorithms available on which we carried out further research:

BUILDING CLASSIFICATION TREES

Construction of a classification tree starts with performing good splits on the data. In this section we define what such a good split is and how we can find such a split. Three impurity measures, re sub situation-error, gini-index and the entropy, for splitting data will be discussed in .The actual splitting and tree construction according to these splits.

IMPURITY MEASURES

A split that will separate the data as much as possible in accordance with the class labels. So the objective is to obtain nodes that contain cases of a single class only as mentioned before. We define impurity as a function of the relative frequencies of the classes in that node's as the relative frequencies of the J different classes in that node To compare all the possible splits of the data you have, a quality of a split as the reduction of impurity that the split achieves must be defined.

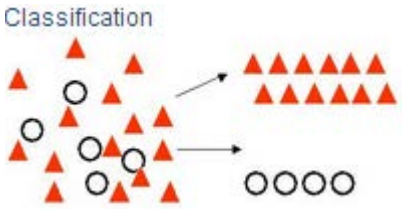
TREE CONSTRUCTION

Building a classification tree starts at the top of the tree with all the data. For all the attributes the best split of the data must be computed. Then the best splits for each of the attributes are compared. The attribute with the best split wins. The split will be executed on the attribute with the best value of the best split (again we consider binary trees). The data is now separated to the corresponding branches and from here the computation on the rest of the nodes will continue in the same manner. Tree construction will finish when there is no more data to separate or no more attributes to separate them by Over fitting and Pruning.

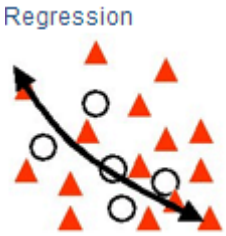
CART ALGORITHM

Decision Trees are commonly used in data mining with the objective of creating a model that predicts the value of a target (or dependent variable) based on the values of several input (or independent variables). The CART or Classification & Regression Trees methodology refer to the following types of decision trees:

Classification Tree: Where is the target variable is categorical and the tree is used to identify the class within which a target variable would likely fall into.



Regression Tree: Where the target variable is continuous and tree is used to predict its values.



CLASSIFICATION TREE

Classification trees are used when for each observation of learning sample we know the class in advance. Classes in learning sample may be provided by user or calculated in accordance with some exogenous rule. For example, for stocks trading project, the class can be computed as a subject to real change of asset price. Let tp be a parent node and tl tr - respectively left and tight child nodes of parent node tp. Consider the learning sample with variable matrix X with M number of variables xj and N observations. Let class vector Y consist of N observations with total amount of K classes. Classification tree is built in accordance with splitting rule - the rule that performs the splitting of learning sample into smaller parts. We already know that each time data have to be divided into two parts with maximum homogeneity:

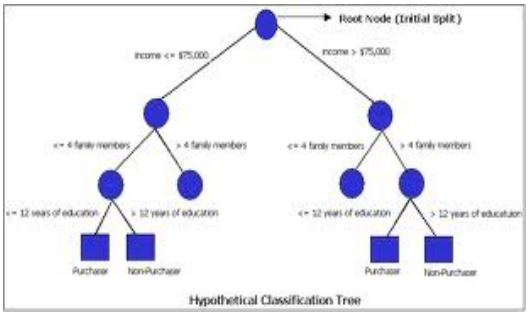


Fig. 5.1 – Hypothetical classification of tree

VI. CONCLUSION & FUTURE WORK

From the calculations we come to the conclusion that the C mean algorithm is an excellent algorithm when we are dealing with classifying a small or medium sized data. It simply provides good performance accuracy every time. A direct algorithm of C-means method requires time proportional to the product of number of patterns and number of clusters identified.

CART algorithm is a decision based algorithm which is used here with C mean algorithm in order to improve the efficiency of classifying data in terms of accuracy, no of clusters formed to classify the data and the time taken to classify the data from an available vast amount of data. We have concluded about the accuracy of proposed algorithm as when the threshold value is 0.2 then we get the accuracy of 74%. There is increment of 0.2 in threshold value and the values of accuracy also change with threshold value. At 0.4 the accuracy is 70%, at 1 the accuracy value is 80% as the last value of threshold 2 we get the final accuracy that is 90% which is better than the previous technique.

VII. FUTUTRE WORK

We can implement the proposed algorithm on the dataset for future work. Future scope of this dissertation is that the classifying technique we have discussed in our dissertation can be subdued using some more powerful technique like Neural network and Genetic Algorithms to classify the data of public interest like.

VIII. REFERENCES

[1] Han, J., Pei, J., and Yin, Y. (2000) Miningfrequent patterns without candidate generation..2000 ACM SIGMOD Intl. Conference on Management of Data.

[2] Leonid Churilov, Adyl Bagirov, Daniel Schwartz, Kate Smith and Michael Dally Drop Out Feature of Student Data for Academic Performance Using Decision Tree, Global Journal of Computer Science and Technology Vol. 10 Issue 2 (Ver 1.0), April 2010.

[3] Fadi Thabtah. A comparison of decision tree classifiers with backpropagation neural networks for multimodal classi_cation problems.Pattern Recognition, 26:953{961, 1993.

[4] S. Bull. Analysis of attitudes toward workplace smoking restrictions. In N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhouse, editors, Case Studies in Biometry, pages 249{271. John Wiley & Sons, New York, NY, 1994.

[5] Shuang Zhang, Xuehua Liu. Rubinfeld.Hedonic prices and the demand for clean air.Journal of Environmental Economics and Management, 5:81{102, 1978.

[6] Shuang Zhang, Xuehua Liu. Penalized discriminant analysis.Annals of Statistics, 23:73{102, 1995.

[7] Vladimir Vapnik. Discriminant analysis by Gaussian mixtures.Journal of theRoyal Statistical Society, Series B, 58:155{176, 1996.

[8] Tao Li, Chengliang Zhang and MitsunoriOgihara. Flexible discriminant analysis by optimal scoring.Journal of the American Statistical Association, 89:1255{1270, 1994.

[9] JiHe , Ah-Hwee Tan and Chew-Lim Tan. Nonparametric Statistical Methods. John Wiley & Sons, New York, NY, 2nd edition, 1999.

[10] Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás. Very simple classification rules perform well on most commonly used datasets. Machine Learning, 11:63{90, 1993.

AUTHORS' PROFILE

Shazia Sultan completed her Bachelor of computer application from Gandhi P.R. college ,Barkatullah university, Bhopal (M.P.) in 2008, and the Master of computer application from BUIT Bhopal (M.P.) in 2011 presently she is pusuing M.Tech degree in Computer Technology and Application from T.I.T college , Rajiv Gandhi proudyogiki vishwavidyalaya, Bhopal (M.P.)

Amar Nayak is currently designated as head of the department in department of Computer Science Engineering, TIT, Bhopal (M.P.), INDIA.