IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 3, March 2016.
www.ijiset.com
ISSN 2348 – 7968

# Implementation of Classification Algorithms and their Comparison for Educational Dataset

**Sharon Carl[1], Glaston D'souza[2] and Linet Varghese[3]**

[1] Information Technology, Xavier Institute of Engineering, Mumbai, Maharashtra 400016, India

[2] Information Technology, Xavier Institute of Engineering, Mumbai, Maharashtra 400016, India

[3] Information Technology, Xavier Institute of Engineering, Mumbai, Maharashtra 400016, India

**Abstract**

Educational data mining concerns with developing methods for discovering knowledge from data from educational datasets.
Data Mining is the analysis step of the KDD, a process of extracting new patterns from large data sets involving methods from statistics and artificial intelligence. In this project, the selected attributes from dataset were applied to Data Mining Algorithms such as Kmeans algorithm, K nearest neighbour algorithm, Decision Tree algorithm, Naïve Bayes algorithm.
In this project we present the comparison of different classification and clustering algorithms using Java. The algorithm used are Kmeans algorithm, K nearest neighbour algorithm, Decision Tree algorithm, Naïve Bayes algorithm.
The Error rates of various Algorithms were compared to bring out the best and effective Algorithm suitable for this dataset.
The aim of this project is how to use suitable data mining algorithms on educational dataset. This paper focuses on comparative analysis of various data mining techniques and algorithms.

## 1. Introduction

The availability of educational data has been growing rapidly, and there is a need to analyze huge amounts of data generated from this educational system. The ability to classify a student's performance is very important in educational environments.
A very promising arena to attain this objective is the use of Data Mining. In fact, one of the most useful Data Mining tasks is classification. Classification is one of the supervised learning techniques that build a model to classify a data item into a predefined class label.
Data mining is the process of automatic classification of cases based on data patterns obtained from a dataset. A number of Algorithms have been developed and implemented to dig out information and discern knowledge patterns that may be constructive for decision support. Once these patterns are extracted they can be used for automatic classification of data. The dataset used in this project is the Educational Dataset.
In this project we are going to compare different data mining techniques for classifying the `attributes in the Educational Dataset. The field of data mining is an emerging research area with important applications in Engineering, Science, Medicine, Business and Education. The size of data base in educational application is large where the number of records in a data set can vary from some thousand to thousands of millions. The size of data is accumulated from different fields exponentially increasing.
The overall aim of this project is to extract information from huge datasets and transform it into understandable structure for further use.
The rest of paper is organized as follows: Section 2 describes problem definition. Section 3 contains proposed system. The Result Analysis is shown in the Section 4. Conclusion is shown in section 5, while References are mentioned in the last section.

## 2. Problem Definition

In this project we are going to implement various data mining algorithms like K means, K Nearest Neighbor, Decision Tree, Naive Bayes etc. on a particular dataset.
The dataset used in this project is the Educational Dataset.
The project aims to use different data mining techniques for classifying the attributes in the Educational Dataset by using different Data Mining Classification Algorithms. Also the project aims to compare their Individual results of the four algorithms to find out the best algorithm in terms of cost and time efficiency. It helps to determine which algorithm is better in terms of output and also to find which algorithms gives the most relevant output for the

ISSN 2348 – 7968

given dataset , so as to give a better and relevant result to the user.
The project is implemented using Java to analyze and compare the various results of the four algorithms used. Java helps in the comparison of those algorithms to bring out the best and effective Algorithm which would give the most relevant output suitable for the Educational dataset.

# 3 Proposed System

## 3.1 System Architecture

This section deals with the architecture of the proposed system model which is shown in Figure 1The subsets of the original dataset are considered for further analysis of Classification Algorithms.

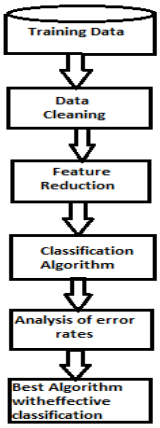


Fig. 1 Architecture of the Proposed system model.

## 3.2 System Analysis

### 3.2.1 Data Selection

The proposed methodology will be used to generate a database for the current study. Before processing of data we will be going to clean the data to remove noise and inconsistency. To remove missing values in the dataset, we will use the cleaning techniques. The experiments and observations will be carried out by using data mining tool i.e. Java. It includes the following phases: Data Cleaning( replacement of missing Values), Data Pre-processing, Feature Reduction (relevant attributes required to perform are selected), Data Mining Classification Algorithms**,** Analysis of error rates produced by Algorithms**,** Identifying the best Algorithm for the dataset.

### 3.2.2 Tool Selection (Language): Java

Java was developed for the purpose of identifying information from raw data gathered from agricultural domains. Data pre-processing, classification, clustering, association, regression and feature selection these standard data mining tasks are supported by JAVA. It is an open source application which is freely available. In JAVA datasets could be in any format. The JAVA will use these automatically even if its not in a particular format. We use JAVA for the classification purpose. A large different number of classifiers are used in JAVA such as Bayes, function, tree etc.
Steps to apply classification techniques on data set and get result in JAVA:
Step 1: Take the input dataset.
Step 2: Apply the classifier algorithm on the whole data set.
Step 3: Note the accuracy given by it and time required for execution.
Step 4: Repeat step 2 and 3 for different classification algorithms on different datasets.
Step 5: Compare the different accuracy provided by the dataset with different classification algorithms and identify the significant classification algorithm for particular dataset .

## 3.2 Methodology

### 3.2.1 Data Mining Techniques

Data mining is a knowledge discovery process to find previously unknown, potentially useful and non-trivial patterns from large repositories of data like the application of data mining techniques to extract knowledge from data. There are interesting relationships discovered among data items .Generally data mining contains several algorithms and techniques for picking out interesting patterns from large data sets .
Data mining techniques are classified into two categories: supervised learning and unsupervised learning.
In supervised learning, a model is built prior to the analysis. We then apply the algorithm to the data in order to estimate the parameters of the model. Classification, Decision Tree, Bayesian Classification, are common examples of supervised learning.
In unsupervised learning, we do not create a model or hypothesis prior to the analysis. We just apply the algorithm directly to the dataset and observe the results. Then a model can be created on the basis of the obtained results.

ISSN 2348 – 7968

Clustering is one of the examples of unsupervised learning. Various data mining techniques such as Classification, Decision Tree, Bayesian Classification, K means and Nearest Neighbour have been used for knowledge discovery from the educational large data sets. Some of the common and useful data mining techniques have been discussed.

### 3.2.1.1 Classification

Classification is a supervised learning technique. It maps the data into predefined groups. It is used to develop a model that can classify the population of records at large level.
Classification algorithm requires that the classes be defined based on the data attribute value. It describes these classes according to the characteristics of the data that is already known to belong to the classes.
The classifier training algorithm uses these pre-defined examples to determine the set of parameters required for proper discrimination. This section discusses some of the useful data mining techniques such as Decision tree, K means, K nearest neighbours, Bayesian Classification etc.

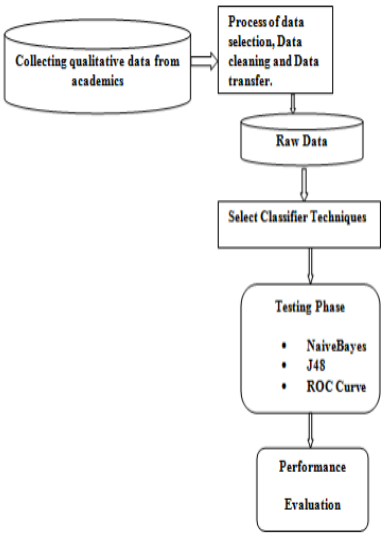


Fig. 2 Methodology of classification technique.

(i) K-means Algorithm :

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into cells. The k-means algorithm is the simplest and most commonly used clustering algorithm employing a square error criterion . It is computationally fast, and iteratively partitions a data set into k disjoint clusters, where the value of k is an algorithmic input. The goal is to obtain the partition (usually of hyper-spherical shape) with the smallest square-error. Suppose k clusters {C1, C2,…, Ck} such that Ck has nk patterns. The mean vector or centre of cluster Ck

$$\mu^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_i^{(k)} \qquad (1)$$

where $n_i$ is number of patterns in cluster $C_i$ , (among exactly k clusters: C1, C2, …, Ck) and x is the point in space representing the given object .

K-Means clustering algorithm:
Select k documents from the collection to form k initial singleton clusters. Estimate k = floor (sqrt(N)).Repeat until termination conditions are satisfied.For every document d, find the cluster i whose centroid is most similar, assign d to cluster i.For every cluster i, re-compute the centroid based on the current member documents.Check for termination -- minimal or no change to the assignment of documents to clusters.Return list of cluster.

The total squared-error :

$$E_K^2 = \sum_{k=1}^{K} e_k^2 \qquad (2)$$

where

$$e_k^2 = \sum_{i=1}^{n_k} (x_i^{(k)} - \mu^k)^T \, (x_i^{(k)} - \mu^k) \qquad (3)$$

(ii) K Nearest Neighbours (KNN) :

The k-nearest neighbour algorithm makes a classification for a given sample without making any assumptions about the distribution of the training and testing data. Each testing sample must be compared to all the samples in the training set in order to classify the sample.
In order to make a decision using this algorithm, the distances between the testing sample and all the samples in the training set must first be calculated. In this any distance measurement may be used .
The Euclidean distance metric requires normalization of all features into the same range. At this point, the k closest neighbours of the given sample are determined where k represents an integer number between 1 and the total number of samples.

ISSN 2348 – 7968

The testing sample is then assigned to the label most frequently represented among the k nearest samples. The value of k that is chosen for this decision rule has an affect on the accuracy of the decision rule The *k-nearest neighbour* classifier is a nonparametric classifier that is said to yield an efficient performance for optimal values of *k*.

Stepwise Implementation:

Step1: Get Cosine Similarity of all the documents.

A o B = x1*x2 + y1*y2

dist(A,0) = sqrt((xa-x0)^2 + (ya-y0)^2) == |A|

Therefore:

sim(A,B) = cos t = A o B/|A|x|B

Step2: Loop

a) Select a Centroid|

b) Check the similarity

c) If sim(A,B) < Threshold

continue;

Step3: Return Clusters .

(iii) Decision Tree Algorithm *:*

A decision tree is a flow chart like tree structure, where each node denotes test on an attribute value, each branch represents the result of the test, and tree leaves represent classes. The drive model can be represented in different forms such as classification (If-Then) rules, decision tree, mathematical formula or neural networks.

Decision tree can easily be converted to classification tree. Decision trees are simple to understand and provide good results even with small data .Decision tree induction algorithms can be used for classification in many application areas, such as Education, Medicine, Manufacturing, Production, Financial analysis, Fraud Detection and Astronomy etc. There are several data mining algorithms such as C4.5, ID3, CART, J48, NB Tree, REP Tree etc. General idea of algorithm Tree structure which has been widely used to represent classification models (a classifier depicted in a flowchart). Decision tree induction algorithms, an inductive learning task use particular facts to make more generalized conclusions. Most decision tree induction algorithms are based on a greedy top- down recursive partitioning strategy for tree growth. They use different variants of impurity measures, like; information gain, gain ratio, and distance-based measures, to select an input attribute to be associated with an internal node. One major drawback of Greedy search is that it usually leads to sub-optimal solutions. A predictive model based on a branching series of Boolean tests, these smaller Boolean tests are less complex than a one-stage classifier. Entropy of decision tree is the information gain measure, is minimized when all values of the target attribute are the same, If we know that commute time will always be short, then entropy = 0. Entropy is maximized when there is an equal chance of all values for the target attribute (the result is random), If commute time = short in 3 instances, medium in 3 instances and long in 3 instances, entropy is maximized .

(iv) Naïve Bayes Algorithm :

Naive Bayes is a simple technique for constructing classifiers which are the models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle that all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. The Naïve Bayes represents a supervised learning method as well as a statistical method for classification. It assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems .

Bayes theorem is as follows: P(H/X)=( P(X/H) . P(H) ) / P(X) .

## 4. Result of analysis of Data mining techniques and algorithm.

This section presents the comparative analysis of different data mining techniques and algorithms which have been used by most of the researchers in educational data mining. A brief summary of these data mining algorithms with their merits and demerits have been discussed. The comparative study of classification algorithms such as Decision Tree, Naïve Bayesian and Neural Networks is shown in table:

| Algorithm | Merits | Demerits |
|---|---|---|
| Decision Tree | It can handle both continuous and discrete data. It provides fast result in classifying unknown | It can't predict the value of a continuous class attribute. It provides error prone results when too many classes are used. Irrelevant attribute affects construction of a decision tree in a bad manner. |

ISSN 2348 – 7968

| | | |
|---|---|---|
| | records.<br><br>It works well with redundant attribute.<br><br>It provides good results with small size tree.<br><br>Results does not affect with outliers.<br><br>It does not require preparation method like normalization.<br><br>It also works well with numeric data. | Small change in data can change the decision tree completely. |
| Naïve Bayesian | It provides high accuracy and speed on large database.<br>It has minimum error rate in comparison to all other classifier.<br>It is easy to understand.<br>It is not sensitive to irreverent features.<br>It handles streaming data well.<br>It can also handle real and discrete values. | It assumes independence of features. So it provides less accuracy. |
| K-Nearest Neighbour | It performs better with missing data.<br>It is easy to implement and debug.<br>It provides more accurate results.<br>Some noise reduction techniques are used that improve the accuracy of classifier. | It requires high calculation complexity.<br>It considers no weight difference between samples.<br>It is sensitive to irrelevant and redundant feature. |
| K-Means | It is Reasonable fast.<br>It is very simple and robust algorithm.<br>It provides best results when data sets are distinct. | It can't work with non-linear data sets.<br>It can't handle noisy data and outliers. |

## 5. Conclusion

The result of this project indicates that capabilities of data mining algorithms on the Educational Dataset. A comparative analysis of various data mining techniques is presented in this project. Many data mining algorithms can be implemented on Educational dataset to predict their future performance.

In this project we have compared the performance and usefulness of different data mining algorithms like K means algorithm, K Nearest Neighbours algorithm, Decision trees algorithm and Naive Bayes algorithm for classification of Educational Dataset.

We have shown that these algorithms improve their classification performance when we apply such algorithms to a dataset. In this way we could measure the performance of the algorithms by comparing their results.

Factors that affect the classifier's performance are: Data set, Number of tuples and attributes, Type of attributes, System configuration.

Thus, a comparative analysis of various data mining techniques is presented in this project.

Due to our analysis on comparison among data mining classification's algorithms (Decision tree, K means, KNN, Bayesian) and analysing the time complexity of the mentioned algorithms we conclude that all K means

ISSN 2348 – 7968

algorithms have less error rate and it is the easier algorithm as compared to KNN and Bayesian.

JAVA is used in this project for classification and comparison. It is the simplest language for classify the data various types. The main aim of this project is to provide a detailed implementation of classification algorithms to the Educational Dataset and their comparison using Java. Every algorithm has their own importance and we use them on the behaviour of the data, but on the basis of this research we found that k-means clustering algorithm is simplest algorithm as compared to other algorithm.

**Acknowledgments**

We would like to thank Fr. Francis D'mello (Director of XIE) for providing us with such an environment so as to achieve goals of our project and supporting us constantly.

We express our sincere gratitude to our Honourable Principal Mr. Y.D.Venkatesh for encouragement and facilities provided to us.

We would like to place on record our deep sense of gratitude to Prof. Chhaya Narvekar, Head of Department Of Information Technology, Xavier Institute of Engineering, Mahim, Mumbai, for her generous guidance help and useful suggestions.

With deep sense of gratitude we acknowledge the guidance of our project guide Prof. Suvarna Bhoir.

The time-to-time assistance and encouragement by her has played an important role in the development of our project.

We would also like to thank our entire Information Technology staff who have willingly co-operated with us in resolving our queries and providing us all the required facilities on time.

## References

[1] Applicability of Clustering and Classification Algorithms for Recruitment Data Mining, International Journal of Computer Applications (0975 – 8887) Volume 4 – No.5, July 2010.

[2] A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data, International Journal of Computer Applications (0975 – 8887) Volume 32– No.8, October 2011.

[3] Comparative Analysis of Classification Algorithms on Different Datasets using weka, International Journal of Computer Applications (0975 – 8887) Volume 54– No.13, September 2012.

[4] Prediction of Higher Education Admissibility using Classification Algorithms, Volume 2, Issue 11, November 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[5] Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation), Computer Engineering and Intelligent Systems www.iiste.org ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online) Vol.4, No.8, 2013.

[6] Weka Approach for Comparative Study of Classification Algorithm, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 4, April 2013.

[7] COMPARISON OF DIFFERENT CLASSIFICATION TECHNIQUES USING DIFFERENT DATASETS, International Journal of Advances in Engineering & Technology, May 2013. ©IJAET.

[8] Comparative Analysis of Data Mining Techniques on Educational Dataset, International Journal of Computer Applications (0975 – 8887) Volume 74– No.5, July 2013.

[9] Comparative Analysis of Bayes and Lazy Classification Algorithms, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 8, August 2013.

[10] "Evaluating Performance of Data Mining Classification Algorithm in Weka, International Journal of Application or Innovation in Engineering & Management (IJAIEM) WebSite:www.ijaiem.org,Volume 2, Issue 10, October 2013.

[11] Comparison the Various Clustering and Classification Algorithms of WEKA Tools, Volume 3, Issue 12, December 2013 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[12] Educational Data mining for Prediction of Student Performance Using Clustering Algorithms, M. Durairaj et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5987-5991

[13] Implementation of Data Mining Techniques to Perform Market Analysis, International Journal of Innovative Research in Computer and Communication Engineering,(An ISO 3297: 2007 Certified Organization)Vol. 2, Issue 11, November 2014.

[14] Predicting Students Final GPA Using Decision Trees: A Case Study, International Journal of Information and Education Technology, Vol. 6, No. 7, July 2015.