COMPARATIVE ANALYSIS OF NAIVE BAYES ALGORITHM AND ID3 ON PREDICTNG STUDENTS'S ACADEMIC PERFORMANCE

[1]I SRILALITA SARWANI, [2]B MEENA
[1]Assistant Professor, [2]Assistant Professor
[1]IT department,
[1]ANITS, Visakhapatnam, India

Abstract: Predicting students' performance is becoming more challenging issue in educational institutions. The existing system does not allow analyzing and monitoring performance of students. The reason behind this is lack of appropriate methods to identify the factors affecting students' performance. Therefore, a systematic review on students' performance by using data mining prediction techniques is proposed. This tool is helpful in improving student performance and has a good impact on students, teachers and educational institutions. We have analyzed the data set containing information about students, such as academic qualifications, marks scored in SSC and Intermediate examinations and rank obtained in entrance examination and end semester results of the previous batch of students. We have proposed a method for predicting student performance by applying the ID3 (Iterative Dichotomise 3) and Naïve Bayes classification algorithms**.**

Keywords -. ID3, Data mining, prediction

I. INTRODUCTION

Data mining is the process of discovering interesting knowledge, such as associations, patterns, changes, significant structures and anomalies, from large amounts of data stored in databases or data warehouses or other information repositories .It has been widely used in recent years due to the availability of huge amounts of data in electronic form, and there is a need for turning such data into useful information and knowledge for large applications. These applications are found in fields such as Artificial Intelligence, Machine Learning, Market Analysis, Statistics and Database Systems, Business Management and Decision Support. Though it is a part of Knowledge Discovery from data but however, in industry, in media, and in the database research milieu, the term data mining is becoming more popular than the longer term of knowledge discovery from data. Data mining consists of about six tasks. These 6 tasks are divided into supervised and unsupervised learning. In Supervised learning a model is build using the available data and describes the attributes of interest which were generally called target attributes in terms of the other remaining attributes. Whereas in the Unsupervised learning no single attribute (target attribute) is taken but relationship among the attributes of the data is established. Classification and the Prediction that comes under supervised machine learning methods. In Classification by constructing the classifier by appropriate techniques the whole data is divided into classes so that when a new value rises it is predicted to fit into any of the classes that is previously generated in the classifier. Prediction is used to predict future values based on the given data.

This work is on the comparative study of some well-known Data Mining techniques namely Decision Trees and Bayesian Networks and classification algorithms namely ID3 Decision Tree generation algorithm and Naïve Bayes algorithm of Bayesian Networks. The results of this paper would provide a tool for college management and teachers to identify bright students and as well as focus on others for their improvement.

2. DATASET USED

The data we used for our work was collected from students. For this, an online application is created in which students have to answer the questionnaire. The questionnaire consists of several questions which focus on student's personal background, parents' education, academic background and interested subjects. All of the attributes were taken into consideration.

The performance of students in their previous examinations was taken into account. Some of the major attributes in our data set are

- Medium of schooling: Used in the context of assessing proficiency in English.

- Marks obtained in SSC: the percentage of marks obtained in SSC was taken.
- Marks obtained in Intermediate
- Rank obtained in entrance:
- End-semester marks of previous exam

. Dataset are divided into four cases of learning datasets each case ranges the learning dataset as follows.

Admitted batch	Number of students
2011	60
2012	59
2013	64
2014	60

Table 1: Tabular representation of date ranges taken in the training dataset

3. METHODOLOGIES

3.1 Decision Trees

A Decision Tree is a data mining technique which is useful in classification. It is a classification scheme which generates a tree and set of rules. The tree and rules represents a model of different classes, for a given dataset. The tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, and each branch represents an outcome of the test. The leaf nodes represent the classes. The top most node in a tree is the root node from which classification begins.

3.2 Iterative Dichotomiser 3 (ID3) Algorithm

J. Ross Quinlan developed ID3(Iterative Dichotomiser 3). ID3 algorithm has several following steps:

- Calculate the entropy of each and every attribute of the data set
- Split the set S i.e set of attributes into subsets using the attribute for which entropy is minimum
- Generate a decision tree node containing that attribute
- Reduce on subsets using remaining attributes.

3.2.1 ID3 Metrics

Entropy

Entropy is an ID3 metric. Entropy of a set S is represented as H(S). It is a measure of the amount of uncertainty in the dataset S.

$$H(S) = -\sum_{xeX} p(x)log2p(x)$$

Where, S - The current dataset for which entropy is being calculated (it changes for every iteration of the ID3 algorithm) X - Set of classes in S

- p(x) - The proportion of the number of elements in class x to the number of elements in set S

When H(S)=0, the given set S is perfectly classified which means all the elements in the set S are of the same class.

In ID3, entropy is calculated for each remaining attribute. The attribute with the smallest entropy is used to split the set S on this iteration. The higher the entropy, the higher the potential to improve the classification.

Information Gain

Information gain IG (A, S) is the measure of the difference in entropy from before and after the set S is split on a given attribute A. In other words, it is the measure on how much uncertainty in set S was reduced after splitting set S on attribute A

$$IG(A,S) = H(S) - \sum_{teT} p(t)H(t)$$

Where,

[1] H(S) - Entropy of set S
[2] T - The subsets created from splitting set S by attribute A such that = ∪
[3] p(t) - The proportion of the number of elements in t to the number of elements in set S
[4] H(t) - Entropy of subset t

In ID3, information gain can be calculated (instead of entropy) for each remaining attribute. The attribute with the largest information gain is used to split the set S on this iteration.

3.3 Bayesian Networks

A Bayesian network of a set of variables X= {X1, X2, …, Xn} represents a joint probability distribution over those variables consisting of a network structure that encodes assertions of conditional independence in the distribution and a set of conditional probability distribution corresponding to that structure. It is graphically represented by directed acyclic graphs(DAG), whose nodes denotes the random variables, which may be observable quantities, latent variables, unknown parameters or hypotheses. Edges represent conditional dependencies, so that nodes which are disconnected represent variables which are conditionally independent of each other.

3.4 Naïve Bayes Algorithm

This is a classification technique based on Bayes theorem with an assumption of independence among predictors. In easy terms, a Naïve Bayes classifier assumes the presence of a particular feature of set of features in a class is unrelated to the presence of any other feature in the class. For example, a fruit may be considered to be an apple if it is red, round, and nearly 3 inches in diameter. Even though if these features depend on each other or on the existence of the other features of the fruit, these properties independently contribute to the probability that the given fruit is an apple and that is why this technique is known to be as 'Naïve'. Models based on these techniques is easy to build and particularly useful for very large datasets. Along with simplicity, this technique is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c).

$$P(c|x) = \frac{P(X|C)P(c)}{P(x)}$$

Where,

P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor

3.4.1 Probabilistic model

Abstractly, naive Bayes is a conditional probabilistic model given a problem instance to be classified, represented by a vector X= (x1, ..., xn) representing some n features (independent variables), it assigns to this instance probability p(Ck | x1,.....,xn)

for each of *K* possible outcomes or classes.

The problem with the above formulation is that if the number of features *n* is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. Using Bayes theorem, the conditional probability can be decomposed as:

$$p(Ck|x) = \frac{p(Ck)\, p(x|Ck)}{p(x)}$$

In plain English, using Bayesian probability terminology, the above equation can be written as:

$$\text{posterior} = \frac{\text{prior X likelihood}}{\text{evidence}}$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on C and the values of the features Fi are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model.

p(Ck , x1,.....,xn)

which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability.

$$p(Ck, x1, \dots\dots, xn) = p(x1, \dots., xn, Ck)$$

$$= p(x1|x2, \dots., xn, Ck)p(x2, \dots., xn, Ck)$$

$$= p(x1|x2, \dots., xn, Ck)p(x2|x3, \dots.xn, Ck)p(x3, \dots., xn, Ck)$$

Now the "naïve" conditional independence assumptions come into play: assume that each feature Fi is conditionally independent of every other feature Fj for j is not equal to i, given the category C. This means that:

$$p(xi|xi+1, \dots., xn, Ck) = p(xi|Ck)$$

for i=1,....,n-1. Thus, the joint model can be expressed as:

$$p(Ck|x1, \dots., xn) \propto p(Ck, x1, \dots., xn)$$

$$\propto p(Ck)\, p(x1|Ck)\, p(x2|Ck)\, p(x3|Ck) \dots..$$

$$\propto p(Ck) \prod_{i=1}^{n} p(xi|Ck)$$

This means that under the above independence assumptions, the conditional distribution over the class variable C is:

$$p(Ck|x1 \dots\dots xn) = \frac{1}{z}\, p(Ck) \prod_{i=1}^{n} p(xi|Ck)$$

where the evidence Z = p(x) is a scaling factor dependent only on x1,.....,xn that is, a constant if the values of the feature variables are known.

4. OBSERVATIONS

A total of 4 cases were applied with increasing in learning dataset as represented in Table 1. One of the case and it's results were described below.

Case 1:

Learning Dataset: admitted year 2011. Total number of entries was 60.

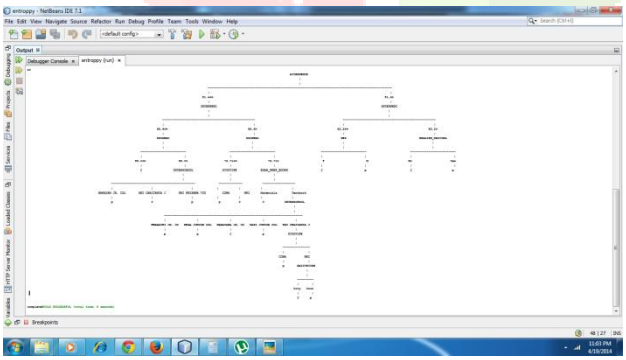
Test Dataset: It consists of 120 students. The responses from their questionnaire are recorded. The result of this method is analyzed to calculate the accuracy of the algorithm and plot the graph.

Observations: some of the predicted values

ORIGINAL VALUES	ID3 PREDICTED VALUES	NAÏVE BAYES PREDICTED VALUES
96.3	80.1355178825	74.750002
96.79	80.1355178825	74.750002
96.66	80.1355178825	74.750002
97.13	80.1355178825	74.750002
99.52	80.1355178825	74.750002
97.39	80.1355178825	74.750002
99.96	80.1355178825	74.750002
98.53	80.1355178825	74.750002
96.96	80.1355178825	74.750002
96.45	80.1355178825	74.750002
97.33	80.1355178825	30.75
92.43	80.1355178825	30.75
99.35	80.1355178825	30.75

Table 2: Results of Case 1

Figure 1: Graphical Representation of Case 1 result



Naïve Bayes Algorithm

- Forecast Accuracy: 70.328
- Algorithm tends to Under-forecast

ID3 Algorithm

- Forecast Accuracy: 82.739
- Algorithm tends to Under-forecast

Results of overall forecast accuracy in each case:

Admitted batch	ID3 ALGORITHM	NAÏVE BAYES ALGORITHM
2011	82.73	70.32
2012	82.55	72.27
2013	94.49	95.44
2014	95.27	97.75

Table 3: Overall forecast accuracies of the algorithms case wise

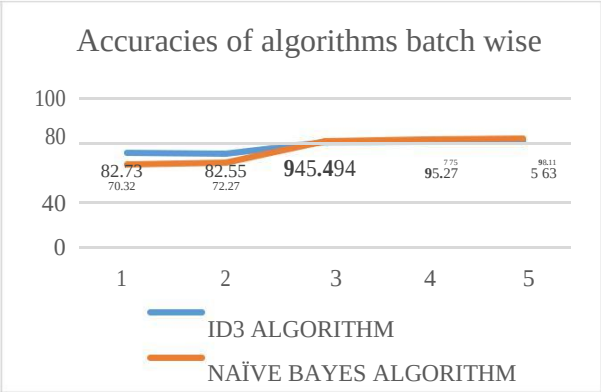


Figure 2 Graph representation of the algorithms accuracies on each case

5. CONCLUSION

In our analysis of applying datasets to the system and the observations from the Table 3 and Figure 2 shown above the Naïve Bayes algorithm has better forecast accuracy and tends to learn faster compared to the ID3 algorithm on predicting students performance on given data set.

6. REFERENCES

[1] Ajay Kumar Pal, Saurabh Pal "Classification Model of Prediction for Placement of students" \.J .Modren Education and Computer Science, 2013,11,49-56.

[2] K. Pal, and S. Pal, "Analysis and Mining of Educational Data for Predicting the Performance of Students", (IJECCE) International Journal of Electronics Communication and Computer Engineering, Vol. 4, Issue

[3] Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification by Tina R. Patil, Mrs. S. S. Sherekar

[4] Data Mining: Concepts and Techniques Second edition by Jiawei Han University of Illinois at Urbana-Champaign Micheline Kamber

[5] Tkinter GUI Application Development HOTSHOT by Bhaskar Chaudhary

[6] Mastering Python for Data Science by S amir madhavan