Implementation of Data Mining Techniques to Classify New Students into Their Classes: A Bayesian Approach

Ramjeet Singh Yadav
Department of Computer Science and Engineering, SET, Sharda University, Greater Noida, UP, India

A.K. Soni
Department of Computer Science and Engineering, SET, Sharda University, Greater Noida, UP, India

Saurabh Pal, Ph.D
Department of MCA, Purvanchal University, Jaunpur, UP, India

ABSTRACT
In educational organizations the classification of new students into appropriate classes is a very challenging task presently. The smartest/intelligent students maybe clustered with the least intelligent in a same class. This problem may be solved by the use of Bayesian classification technique which considers the academic achievements of the students. In present research an attempt has been made to explore Bayesian classification to solve the allocation problem of new students. Based on the present study it is suggested that performance of Bayesian classification technique is more suitable compared to rest of techniques such as genetic algorithm method.

General Terms
Pattern Recognition, Bayesian Classification, Prior Probability.

Keywords
Classification, Bayesian Classification, Prior Probability, New Student Allocation Problem

1. INTRODUCTION
Clustering of newly registered large number of students is a new problem in Indian educational domain. Lack of clustering of new student needs serious handling to avoid other educational problem resulting by classes having both smart and stupid students. Recently, Mankad have reported an evolving rule based model for identification of multiple intelligence [1]. Their genetic-fuzzy hybrid model identifies human intelligence. Zukhri and Omar have reported successful application of Genetic Algorithm for solving difficult optimization problems in new students' allocation problem [2].

General clustering area is well documented, but clustering of new students is highly lacking in the universities. Many researches involve application of GA. In statistics the popular approach to solve clustering problem is agglomerative method [3] which has few demerits especially related with very large objects [4]. Students' allocation problem is constrained by multi-dimensional bin packing [5]. This can be applied, if objective is to minimize the number of classes. Since the objective is to minimize the gap of intelligence in each class, student allocation should be viewed as clustering rather than bin packing. Susanto used Fuzzy C-Means algorithm (FCM) to solve this problem [6]. They clustered students of various subjects based on their scores of prerequisite subjects. This good work also lacks advantage of FCM because it involves 20 students only. Statistical approaches like Agglomerative Methods (AM) can solve clustering [4]. But this cannot apply directly and need modification. The performance of modification depends on the data distribution. In normal distribution, it generates classes with the largest intelligence gap [7]. Therefore, the first class has the lowest gap, and the last class has the highest gap. To avoid problems that can be happened as a result of the random method, universities cluster their new student based on the ranking of academic score (i.e. subject scores, IQ, etc.). In such clustering, the first cluster of students at top level will be clustered as the first class. The next level will be clustered as the next class, and so on. With this method, the universities may expect that there will be clusters of students in each class having comparable intelligence. Suppose there are 25 students and 5 classes, the sorting ways may be N (25, 5) = 2,436,684,974,110,751. Such very large size classification by traditional methods can achieve the local optimum solution only. However, Bayesian classification technique has potential to solve this problem in fairer manner.

The Partition Based Chromosomal Representation (PBCR) approach is least suitable because it search wide space in place of real problem which depends on the number of students (width of chromosome) without considering the number of classes [8]. New student allocation problem can be solved by Bayesian Classification technique [9]. This representation is very simple in comparison to chromosome involving the distribution of new students in each class directly.

2. BAYESIAN CLASSIFICATION
Bayesian, the statistical classifiers, can predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Bayesian classifiers assume an attribute value on a given class which is independent of the values of the other attributes. It is generally known as class conditional independence. Bayesian classification based on Bayes' theorem is described below:

2.1. Bayes' Theorem
Consider X as data tuple. In Bayesian terms, X is considered evidence. It is described by measurements made on a set of n attributes. Let H be some hypothesis, such as that the data tuple X belongs to a specified class C. For classification problems, there is need to determine $P(H/X)$, the probability that the hypothesis H holds given the evidence or observed data tuple X. In other words, there is consideration of probability that tuple X belongs to class C, given that the attribute description of X are known. $P(H/X)$ is the posterior probability of H conditioned on X and $P(H)$ is the prior probability of H. The Bayes' theorem is given below:

$$P(H/X) = \frac{P(X/H)P(H)}{P(X)}$$

(1)

The Bayesian classifier works as follows:

1. Let *D* be a training set of tuples and their class labels. Each tuple is represented by *n*-dimensional attributes vector, $X = (x_1, x_2, \dots, x_n)$, depicting n measurements made on the tuple from *n* attributes, respectively, $A_1, A_2, \dots, A_n$.
2. Suppose, there are *m* classes, $C_1, C_2, \dots, C_m$. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X belongs to the class C_i if and only if $P(C_i/X) > P(C_j/X)$

for $1 \leq j \leq m, j \neq i$. Thus we maximize $P(C_i/X)$. The class C_i for which $P(C_i/X)$ is maximized is called maximum posterior hypothesis.

3. As $P(X)$ is constant for all classes, only $P(X/C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \cdots = P(C_m)$, and there will be maximization of $P(X/C_i)$. Otherwise, maximization will be $P(X/C_i)P(C_i)$.
4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X/C_i)$. In order to reduce computation in evaluate $P(X/C_i)$, the Naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of tuple. Thus,

$$P(X/C_i) = \prod_{k=1}^{n} P(x_k/C_i) \quad (2)$$

5. In order to predict the class label of X, $P(X/C_i)\,P(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of tuple X is the class C_i if and only if $P(X/CiPCi > PX/CjPCj$ for $1 \leq j \leq m, j \neq i$. In other words, they predict class label is the class C_i for which $P(X/C_i)P(C_i)$ is the maximum.

2.2. Laplacian Correction

The Laplacian correction deals with zero probability values. Recall the estimation $P(X/C_i) = \prod_{k=1}^{n} P(x_k/C_i)$ based on the class independence assumption. What if there is a class, C_i and X has an attribute value x_k, such that none of the samples in C_i has that attribute value? In the case $P(x_k/C_i) = 0$, which results in $P(x_k/C_i) = 0$ even though $P(x_k/C_i)$ for all the other attributes in X may be large. There is a simple trick to avoid this problem. Assumption that our training set is so large that adding one to each count would only make a negligible difference in the estimated probabilities, yet would avoid the case of zero probability values. This technique is known as Laplacian correction.

3. EXPERIMENTAL RESULTS

The predicted class label of students using Bayesian classification with the help of training data is given in Table 1. There are 14 data sets belonging to the class first, 14 data sets belonging to class second, 14 data sets belonging to class third and 8 data sets belonging to class fail. The training data sets are described by attributes: end semester marks, class test grade, seminar performance, assignment, general proficiency, attendance, and lab work. The class label attribute and class type, have four distinct values namely first, second, third and fail. The allocation of any new student is shown in Table 2:

Table 1: Class-labeled training tuples from the Students Data Set

S.No.	End Semester Marks	Class Test Grade	Seminar Performance	Assignment	General Proficiency	Attendance	Lab Work	Class Type
1.	FIRST	GOOD	GOOD	YES	YES	GOOD	YES	FIST
2.	FIRST	GOOD	AVERAGE	YES	NO	GOOD	YES	FIST
3.	FIRST	GOOD	AVERAGE	NO	NO	AVERGAE	NO	FIST
4.	FIRST	AVERAGE	GOOD	NO	NO	GOOD	YES	FIST
5.	FIRST	AVERAGE	AVERAGE	NO	YES	GOOD	YES	FIST
6.	FIRST	POOR	AVERAGE	NO	NO	AVERAGE	YES	FIST
7.	FIRST	POOR	AVERAGE	NO	NO	POOR	NO	SECOND
8.	FIRST	AVERAGE	POOR	YES	YES	AVERAGE	NO	FIRST
9.	FIRST	POOR	POOR	NO	NO	POOE	NO	THIRD
10.	FIRST	AVERAGE	AVERAGE	YES	YES	GOOD	NO	FIRST
11.	SECOND	GOOD	GOOD	YES	YES	GOOD	YES	FIRST
12.	SECOND	GOOD	AVERAGE	YES	YES	GOOD	YES	FIRST
13.	SECOND	GOOD	AVERAGE	YES	NO	GOOD	NO	FIRST
14.	SECOND	AVERAGE	GOOD	YES	YES	GOOD	NO	FIRST
15.	SECOND	GOOD	AVERAGE	YES	YES	AVERAGE	YES	FIRST
16.	SECOND	GOOD	AVERAGE	YES	YES	POOR	YES	SECOND
17.	SECOND	AVERAGE	AVERAGE	YES	YES	GOOD	YES	SECOND
18.	SECOND	AVERAGE	AVERAGE	YES	YES	POOR	YES	SECOND
19.	SECOND	POOR	AVERAGE	NO	YES	GOOD	YES	SECOND
20.	SECOND	AVERAGE	POOR	YES	NO	AVERAGE	YES	SECOND
21.	SECOND	POOR	AVERAGE	NO	YES	POOR	NO	THIRD
22.	SECOND	POOR	POOR	YES	YES	AVERAGE	YES	THIRD
23.	SECOND	POOR	POOR	NO	NO	AVERAGE	YES	THIRD
24.	SECOND	POOR	POOR	YES	YES	GOOD	YES	THIRD
25.	SECOND	POOR	POOR	YES	YES	POOR	YES	THIRD
26.	SECOND	POOR	POOR	NO	NO	POOR	YES	FAIL
27.	THIRD	GOOD	GOOD	YES	YES	GOOD	YES	FIRST
28.	THIRD	AVERAGE	GOOD	YES	YES	GOOD	YES	SECOND
29.	THIRD	GOOD	AVERAGE	YES	YES	GOOD	YES	SECOND
30.	THIRD	GOOD	GOOD	YES	YES	AVERAGE	YES	SECOND
31.	THIRD	GOOD	GOOD	NO	NO	GOOD	YES	SECOND
32.	THIRD	AVERAGE	AVERAGE	YES	YES	GOOD	YES	SECOND
33.	THIRD	AVERAGE	AVERAGE	NO	YES	AVERAGE	YES	THIRD
34.	THIRD	AVERAGE	GOOD	NO	NO	GOOD	YES	THIRD

35.	THIRD	GOOD	AVERAGE	NO	YES	AVERAGE	YES	THIRD
36.	THIRD	AVERAGE	POOR	NO	NO	AVERAGE	YES	THIRD
37.	THIRD	POOR	AVERAGE	YES	NO	AVERAGE	YES	THIRD
38.	THIRD	POOR	AVERAGE	NO	YES	POOR	YES	FAIL
39.	THIRD	AVERAGE	AVERAGE	NO	YES	POOR	YES	THIRD
40.	THIRD	POOR	POOR	NO	NO	GOOD	NO	THIRD
41.	THIRD	POOR	POOR	NO	YES	POOR	YES	FAIL
42.	THIRD	POOR	POOR	NO	NO	POOR	NO	FAIL
43.	FAIL	GOOD	GOOD	YES	YES	GOOD	YES	SECOND
44.	FAIL	GOOD	GOOD	YES	YES	AVERAGE	YES	SECOND
45.	FAIL	AVERAGE	GOOD	YES	YES	AVERAGE	YES	THIRD
46.	FAIL	POOR	POOR	YES	YES	AVERAGE	NO	FAIL
47.	FAIL	GOOD	POOR	NO	YES	POOR	YES	FAIL
48.	FAIL	POOR	POOR	NO	NO	POOR	YES	FAIL
49.	FAIL	AVERAGE	AVERAGE	YES	YES	GOOD	YES	SECOND
50.	FAIL	POOR	GOOD	NO	NO	POOR	NO	FAIL

Table 2: Data set for a new Student

S.No.	End Semester Marks	Class Test Grade	Seminar Performance	Assignment	General Proficiency	Attendance	Lab Work
1.	FIRST	GOOD	GOOD	YES	NO	GOOD	YES

The need to maximize $P(X/C_i)$, for $i = 1,2,3,4$, $P(C_i)$, the prior probability of each class, can be computed based on the training data set.

$$P(Class\ Type = First) = \frac{14}{50} = 0.2800$$

$$P(Class\ Type = Second) = \frac{14}{50} = 0.280$$

$$P(Class\ Type = Third) = \frac{14}{50} = 0.2800$$

$$P(Class\ Type = First) = \frac{14}{50} = 0.1600$$

To compute $P(X/C_i)$ for $i = 1,2,3,4$, $P(C_i)$, we compute the following probabilities:

$$P(End\ Semester\ Marks = First)/Class\ Type = First) = \frac{9}{18} = 0.5000$$

$$P(End\ Semester\ Marks = First)/Class\ Type = Second) = \frac{2}{18} = 0.1111$$

$$P(End\ Semester\ Marks = First)/Class\ Type = Third) = \frac{2}{18} = 0.1111$$

$$P(End\ Semester\ Marks = First)/Class\ Type = Fail) = \frac{1}{12} = 0.0833$$

$$P(Class\ Test = Good)/Class\ Type = First) = \frac{8}{14} = 0.5714$$

$$P(Class\ Test = Good)/Class\ Type = Second) = \frac{6}{14} = 0.4286$$

$$P(Class\ Test = Good)/Class\ Type = Third) = \frac{1}{14} = 0.0714$$

$$P(Class\ Test = Good)/Class\ Type = Fail) = \frac{1}{8} = 0.125$$

$$P(Seminar\ Performance = Good)/Class\ Type = First) = \frac{5}{14} = 0.3571$$

$$P(Seminar\ Performance = Good)/Class\ Type = Second) = \frac{5}{14} = 0.3571$$

$$P(Seminar\ Performance = Good)/Class\ Type = Third) = \frac{2}{14} = 0.1429$$

$$P(Seminar\ Performance = Good)/Class\ Type = Fail) = \frac{1}{8} = 0.1250$$

$$P(Assignment = Yes)/Class\ Type = First) = \frac{10}{14} = 0.7143$$

$$P(Assignment = Yes)/Class\ Type = Second) = \frac{11}{14} = 0.7857$$

$$P(Assignment = Yes)/Class\ Type = Third) = \frac{5}{14} = 0.3571$$

$$P(Assignment = Yes)/Class\ Type = Fail) = \frac{1}{8} = 0.1250$$

$$P(General\ Proeficiency = No)/Class\ Type = First) = \frac{5}{14} = 0.3571$$

$$P(General\ Proeficiency = No)/Class\ Type = Second) = \frac{3}{14} = 0.2143$$

$$P(General\ Proeficiency = No)/Class\ Type = Third) = \frac{6}{14} = 0.4286$$

$$P(General\ Proeficiency = No)/Class\ Type = Fail) = \frac{4}{8} = 0.5000$$

$$P(Attendance = Good)/Class\ Type = First) = \frac{11}{17} = 0.6471$$

$$P(Attendance = Good)/Class\ Type = Second) = \frac{9}{17} = 0.5294$$

$$P(Attendance = Good)/Class\ Type = Third) = \frac{4}{17} = 0.2353$$

$$P(Attendance = Good)/Class\ Type = Fail) = \frac{1}{11} = 0.9091$$

$$P(Lab\ Work = Yes)/Class\ Type = First) = \frac{9}{14} = 0.6429$$

$$P(Lab\ Work = Yes)/Class\ Type = Second) = \frac{13}{14} = 0.9296$$

$$P(Lab\ Work = Yes)/Class\ Type = Third) = \frac{11}{14} = 0.7857$$

$$P(Lab\ Work = Yes)/Class\ Type = Fail) = \frac{5}{8} = 0.3571$$

Using above probabilities, we obtain:

P(New Student/Class Type = First) = P(End Semester Marks = First/Class Type=First)× P(Class Test = Good/Class Type=First) × P(Seminar Performance = Good/Class Type = First) × P(Assignment = Yes/Class Type = First) × P(General Proficiency = No/Class Type=First)× P(Attendance = Good/Class Type = First) × P(Lab work = Yes/Class Type = First)

= 05×0.5714×0.3571×0.7174×0.3571×0.6471× 0.6429*= 0.01087*

Similarly, we can find out

P(New Student/Class Type = Second)= 0.1111 ×0.4286×0.3570×0.7857×0.2143×0.5294×0.9296= 0.00140

P(New Student/Class Type = Third)= 0.1111 ×0.0714×0.1429×0.3571×0.4286×0.2353×0.7857= 0.000032

P(New Student/Class Type = Fail)= 0.0625 ×0.125×0.125×0.125×0.5×0.0.0833×0.3571= 0.000001816

To find the class C_i, that maximize $P(X/C_i)P(C_i)$, we compute

P(New Student/Class Type = First)P(Class Type = First)=0.1087×0.28 = 0.0030436

P(New Student/Class Type = Second)P(Class Type = Second)= 0.00140×0.28 = 0.000392

P(New Student/Class Type = Third)P(Class Type = Third)= 0.000032×0.28 = 0.00000896

*P(New Student/Class Type = Fail)P(Class Type = Fail)= 0.*000001816*×0.16 = 0.0000016*

In this way the Bayesian classifier reliably predicts the new student belonging to class first. In the same manner another new student can be fitted to their respective class based on the performance.

4. CONCLUSION AND FUTURE WORK

It is evident that the Bayesian classification technique, used for the first time in the present work for educational institute, is best model for solving new student allocation problem. This statistical based Bayesian classification technique and associated methods have also been implemented and tested with expected results. The proposed technique may serve as a potential benchmark to monitor the progression of students modeling in educational domain. In addition to this it will also enhances the decision making of academic planners via improvement in the future academic results.

It future research the combination of technique of Subtractive clustering technique, Genetic Algorithm and Artificial Neural Network techniques (i.e. hybrid Fuzzy Expert system) are required urgently in relation to evaluation of both students and teachers academic performance. Such practice may develop adaptive learning system and Intelligent Tutoring System for Internet based education like Distance Education and instructional design.

ACKNOWLEDGEMENTS

I would like to express my deep sense of gratitude and respect to late Prof. Pervez Ahmed for his excellent guidance and suggestions provided to me during this work.

REFERENCES

[1] Mankad, K., Sajja, P.S., and Akerkar, R. (2011). Evolving Rules Using Genetic Fuzzy Approach: An Educational Case Study. International Journal on Soft Computing, 2(1), 35-46.

[2] Zukhri, Z., and Omar, K. (2008). Genetic algorithm with center based Chromosomal representation to solve new student allocation problem. Media Informatika, 5(2), 79-86.

[3] Everitt, B.S., Landau, S. and Leese, M. (2001). Cluster Analysis. London: Heinemann Educational Books Ltd.

[4] Cole, R.M. (1998). Clustering with Genetic Algorithms. Master Thesis University of Western Australia.

[5] Wright, M. (2001). Experiments with a Plateau-Rich Solution Space. Proceedings of the 4th Meta Heuristics International Conference, 317-320.

[6] Susanto, S., Suharto, I. and Sukapto, P. (2002). Using Fuzzy Clustering Algorithm for Allocation of Students." Transaction on Engineering and Technology Education, 1(2), 245-248.

[7] Zukhri, Z. and Omar, K. (2006). Modification of Agglomerative Methods to Cluster New Students into Their Classes. Proceedings of the 1st International Conference on Mathematics and Statistics, 493-498.

[8] Zukhri, Z., and Omar, K. (2007). Comparative Evaluation of Genetic Algorithm and Modification of Agglomerative Method in New Students Allocation Problem. Proceedings of Application of Information Technology National Seminar, B9-B12.

[9] Zukhri, Z., and Omar, K. (2006). Implementation of Genetic Algorithms to Cluster New Students into Their Classes. Seminar Nasional Aplikasi Teknologi Informasi 2006 (SNATI 2006), 101-103.