Classification of Students Using psychometric tests with the help of Incremental Naive Bayes Algorithm

Roshani Ade
Research Scholar, Sipna College of Engineering and Technology, Amravati University, Amravati, India

P. R. Deshmukh, PhD
Professor, Sipna College of Engineering and Technology, Amravati University, Amravati, India

ABSTRACT
In this study, we validate that the incremental leaning as a technique in data mining can be used to classify the students according to their interest by conducting some aptitude test including psychometric tests on students. So that the students can get the correct carrier choice, student can learn the subject in which he/she is interested and improve their as well as institutes performance in terms of result.

Recent years have observed very increasing interest in the topic of incremental learning, as it is having the ability to learn from new data introduces with the system even after the classifier has been produced from the formerly available data. It is required that the leaning should be done without accessing previously learned data and must remember previously acquired knowledge. This can be achieved by using incremental naïve bayes classifier.

General Terms
Incremental Learning, Naïve Bayes Algorithm

Keywords
Incremental Naïve Bayes

1. INTRODUCTION
Now a day's it has been proved that, there is a psychological impact on choosing carrier of a student. It can be proved that by conducting various kinds of tests including personality test, interest test, aptitude test, students can be classified according to their interest and can have right career choice at the beginning of their career. So that, they can study the subject of their interest and improve their performance. In the literature [1], the classification of educational background of students have been done by using musical intelligence and prediction was done with the help of genetic neural networks. In [2], the performance of students in computer programming course according to learning style was classified. The performance of students was predicted in the distance education [3].

The amount of students data in the databases is growing day by day, so the knowledge taken out from these data need to be updated continuously. The problem with the non-incremental techniques is that, it requires high computational effort. Incremental learning is very much important for handling large amount of data because of two reasons. First is , it is not possible to collect all the data beforehand before training and the other is modifying the existing system is much better in terms of time and cost as compared to building totally new system.

The Naïve bayes classifier is the classifier which confines the assumption that every attribute is independent of all the other attribute[4]. The unique characteristics of naïve bayes are, they are robust to isolate noise points since such points are averaged out when approximating conditional probabilities of data[5,6,7]. It also handles missing values just by ignoring the examples during classifier building. Correlated attributes can degrade the performance of naïve bayes classifiers because the conditional independence assumption no longer holds for such attributes[8,9]. If the irrelevant attributes, then they are robust. If Xi is and irrelevant attribute, then P(Xi|Y) becomes almost uniformly distributed. The conditional probability of Xi has no impact on the on the whole computation of the posterior probability.

1.1 Incremental Learning
Even if the current work in the incremental learning from the knowledge finding and data mining perspective, many new algorithms have been developed and applied to the various field[10]. For example, IPCA for incremental learning using PCA was proposed in [11] for online pattern classification to handle the chunk of samples at the same time. The Genetic Algorithm concept was used as a base algorithm for ILGA can supply the different initialization situation [12].
In [13] and [14], the incremental leaning for independent navigation system was proposed. Some additional work on incremental leaning includes, the parameter incremental learning algorithm for Multilayer Perceptron[15], for concept drift detection [16,17], online face recognition [18, 19].

Section 2 gives the methodology applied in the work and the description of the algorithm used, while section 3 mentions the dataset used and section 4 talks about the result and finally section 5 concludes the research.

2. METHODOLOGY
The naïve bayes classifier estimates the class conditional probability by considering that the attributes are conditionally independent.

The conditional independent assumption is stated as

$$P(X|Y=y)=\prod_{k=1}^{n}P(X_k|Y=y)$$

Where, each attribute set $X=(X_1, X_2,\dots ,X_n)$

For conditional independence, le X, Y and Z are three sets of random variables. The variables in X is conditionally independent of Y, if the Z is given, the condition below holds.

$$P(X|Y,Z)=P(X|Z)$$

To classify the test record, the naïve bayes classifier computes the posterior probability for each class Y

$$P(Y|X)=\frac{P(Y)\prod_{k=1}^{n}P(X_k|Y)}{P(X)}$$

P(X) is fixed for every Y, it is enough to choose the class that maximizes the numerator term, $\prod_{k=1}^{n} P(X_k|Y)$, there are different approached for estimating conditional probability for different attribute types.

2.1 Algorithm Description

For Training the data:

1. Initialize all values=0, total=0, consider each one training sample at a time.
2. For each training sample, increment the vector value count.
3. These counts and total are converted into probabilities by calculating the probability of each vector x in each class y using P(x|y)
4. Prior probabilities are calculated as the portion of all training samples which are in class y.

For Testing the data:

1. Get the test samples
2. Probability is calculated using P(y) times the calculated probability of each vector in class y.
3. If the probability of the nearly all possible class is at least two times the probability of the subsequently possible class then the decisions is of Naïve Bayes else the class with the maximum probability is the last decision.

3. DATASET USED

The dataset is created by conducting psychometric test on 250 students of age group 16 to 20. The dataset contains 250 samples,10 attributes and 7 classes.

The attributes are shown in Table 1. All the attributes are numeric attributes. The Fig 1 shows the histogram of all the attributes, total score and the class.

Table 1 : The attributes in the dataset and the score criteria

Actors	**A (Self Awareness)**	**B (Empathy)**	**C (Self Motivation)**	**D (Emotional Stability)**	**E (Managing Relations)**	**F (Integrity)**	**G (Self Development)**	**H (Value Orientation)**	**I (Commitment)**	**J (Altruistic Behavior)**
High	11and above	15 and above	18 and above	11 and above	12 and above	8 and above	6 and above	6 and above	6 and above	6and above
Normal	4 to 10	7 to 14	9 to 17	4 to 10	5 to 11	4 to 7	2 to 5	2 to 5	2 to 5	2 to 5
Low	3 and below	6 and below	8 and below	3 and below	4 and below	3 and below	1 and below	1 and below	1 and below	1 and below

4. RESULTS AND PERFORMANCE EVALUATION

In this section, the experimental result and the statistical characteristics of the dataset are discussed. Table 2, shows the statistical characteristics of every attribute in the dataset.

The performance of a model is expressed in Table 3, in terms of kappa statistics and the error rate which can be calculated by using below equations**.**

$$\text{kappa statistics} = \frac{\text{TA} - \text{RA}}{1 - \text{RA}}$$

$$\text{TA} = \frac{\text{TP} + \text{TN}}{\text{TTP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{RA} = \frac{(\text{TN} + \text{FP}) * (\text{TN} + \text{FP}) + (\text{FN} + \text{TP}) * (\text{FP} + \text{TP})}{\text{Total} * \text{Total}}$$

Where, TA : Total Accuracy, RA: Random Accuracy, TP : True Positive, TN : True Negative, FP: False Positive, FN: False Negative.

The error rate is calculated by using the equation.

$$\text{Error rate} = \frac{\text{number of wrong predictions}}{\text{Total number of preditions}}$$

Table 3 shows the comparative results of Multilayer Perceptron and the incremental naïve bayes algorithm in terms of mentioned performance parameters, which shows that Incremental Naïve Bayes gives good results with less time as compared to Multilayer Perceptron. The performance accuracy by class by using incremental Naïve Bayes is shown in Table 4 and confusion matrix of the classified instances is shown in Table 5.

Fig. 1: Histrogram of all attributes of the students dataset

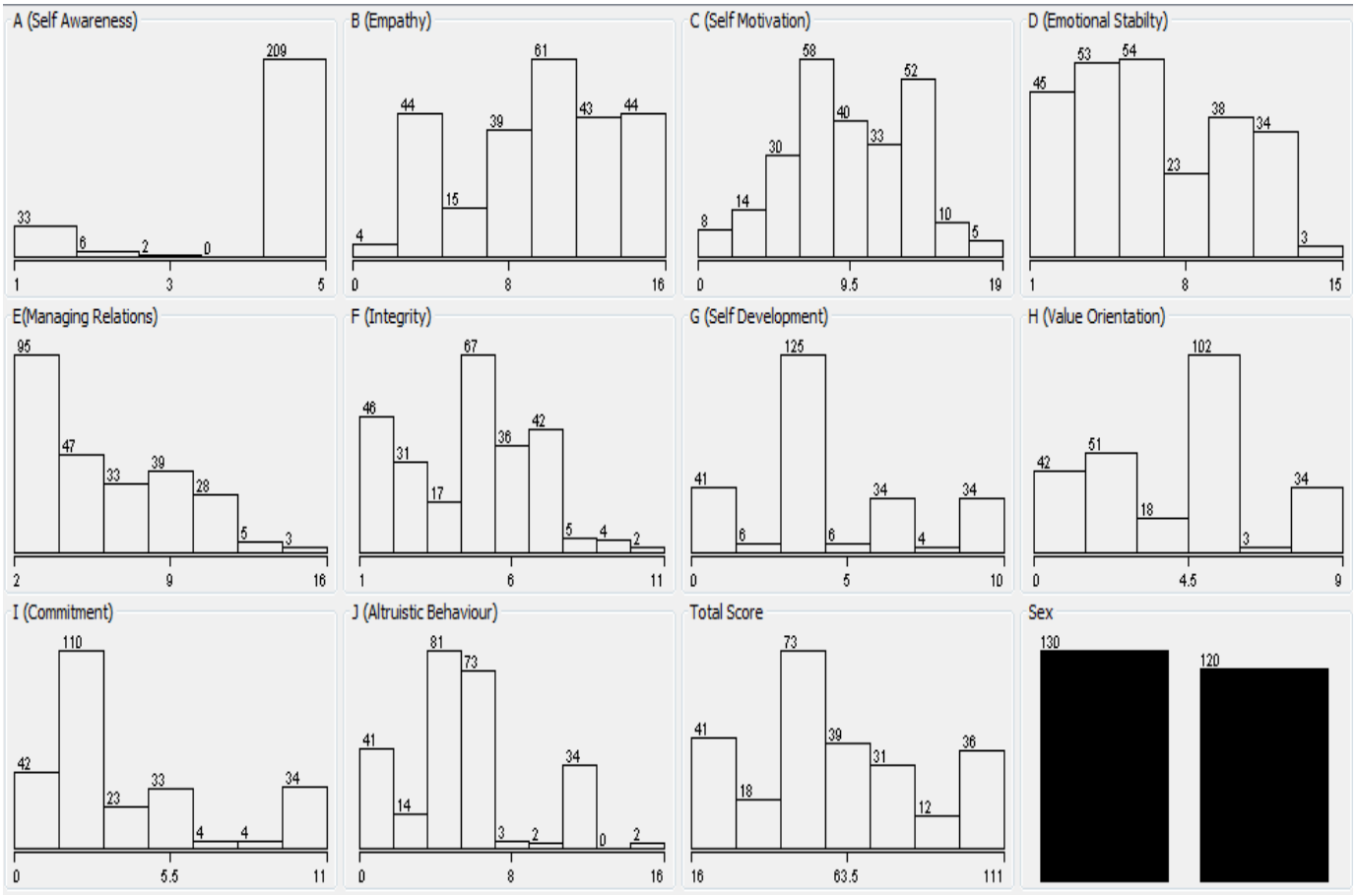

Table 2: Mean and Standard Deviation of the attributes in the dataset

Attributes	Class						
	1	2	3	4	5	6	7
	0.2	0.23	0.14	0.1	0.1	0.07	0.16
Mean (A)	5.3333	5.3333	5.3333	5.3333	5.3333	5.3333	1.5935
Std. Dev. (A)	0.2222	0.2222	0.2222	0.2222	0.2222	0.2222	0.5284
Mean(B)	8.8039	10.9825	6.9143	11.96	14.4167	14.7647	3.5366
Std. Dev.(B)	2.5051	1.9779	2.8921	0.5987	1.3514	1.0017	0.8861
Mean(C)	7.3061	11.463	7.6603	12.0333	13.7222	14.5915	4.5054
Std. Dev.(C)	0.5854	2.1609	2.7565	0.9902	1.669	1.7824	1.4564
Mean(D)	3.9608	7.9825	6.6571	11.12	11.7917	11.9412	1.7805
Std. Dev.(D)	0.1941	1.5838	1.0404	0.8635	1.0793	0.8022	0.4139
Mean(E)	2	7.5263	4.8286	11.08	10.5	11.4118	2.9756
Std. Dev.(E)	0.1667	1.6869	0.6963	1.6473	1.472	1.9421	0.1667
Mean(F)	5.0588	5.2456	2.8286	6.36	6.875	7.5882	2.0488
Std. Dev.(F)	0.3658	1.2605	0.6088	0.8429	0.7253	1.2861	0.3085
Mean(G)	3.0392	3.8772	3.0286	6.2	8.5417	8.7059	0.9756
Std. Dev.(G)	0.1941	1.2435	0.1667	0.6928	1.2576	0.6655	0.1667
Mean(H)	5	4.0526	2.0857	6.08	8.375	8.4706	0.9756
Std. Dev.(H)	0.1667	1.1305	0.4389	0.2713	1.4948	1.0357	0.1667
Mean(I)	3	3.4912	2.9714	6.04	9.4583	9.8824	0.9756
Std. Dev.(I)	0.1667	1.2443	0.4463	0.196	1.2903	0.5823	0.1667
Mean(J)	5.816	4.0594	3.6923	6.4985	10.8718	10.6425	1.2008
Std. Dev.(J)	1.0072	1.004	0.2051	1.0161	2.0307	1.029	0.2051

Table 3: Error Calculation of Multilayer Perceptron and Incremental Naïve Bayes Classifiers

Performance Parameters	Multilayer Perceptron	Incremental Naïve Bayes
Correctly Classified Instances	88 %	89.6 %
Incorrectly Classified Instances	12 %	10.4 %
Kappa statistic	0.8566	0.8759
Mean absolute error	0.0406	0.0306
Root mean squared error	0.1528	0.1591
Relative absolute error	16.9692 %	12.7845 %
Root relative squared error	44.2096 %	46.008 %
Time Taken	3.9 Sec	Less than 1 Sec

Table 4: Performance accuracy by class of Incremental Naïve Bayes

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.98	0.005	0.98	0.98	0.98	1	1
0.982	0.005	0.982	0.982	0.982	0.999	2
0.971	0	1	0.971	0.986	0.999	3
0.88	0.009	0.917	0.88	0.898	0.982	4
0.292	0.027	0.538	0.292	0.378	0.949	5
0.824	0.069	0.467	0.824	0.596	0.956	6
1	0	1	1	1	1	7
0.896	0.01	0.903	0.896	0.893	0.99	Weighted Average

Table 5: Confusion Matrix by using Incremental Naïve Bayes

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
Class 1	50	1	0	0	0	0	0
Class 2	0	56	0	1	0	0	0
Class 3	1	0	34	0	0	0	0
Class 4	0	0	0	22	3	0	0
Class 5	0	0	0	1	7	16	0
Class 6	0	0	0	0	3	14	0
Class 7	0	0	0	0	0	0	41

5. CONCLUSION

The naïve bayes incremental technique can be used to classify the students according to their interest without retraining the system from scratch. The comparative result of multilayer perception and the naïve bayes incremental leaning shows that naïve bayes incremental technique gives good result in less time. The future scope of this study is that, the mapping functions can be designed with the help of different classifiers which can map the knowledge between previous data and the newly coming data, the hybrid features of a students, including physical fitness and the test results can be uses for the classification of a students, The naïve bayes algorithm can be used as a weak classifier in the ensemble concept for incremental learning.

6. REFERENCES

[1] Firat Hardlac, "Classification of educational backgrounds of students using musical intelligence and perception with the help of genetic neural networks", Expert systesm with applications vol. 36, 2009, pp.6708-6713.

[2] N. M. Norwawai, S. F. Abdusalam, C. F. Hibadulla, B. M. Shuaibu, "Classsification of students performance in computer programming course according to learning style", 2nd conference on data mining and optimization, 27-28 Oct. 2009, Selangor, Malaysia.

[3] S. Kotsiantis, K. Patriarcheas, M. Xenos, "A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education", Knowledge-Based Systems vol 23, 2010 pp. 529–535.

[4] Remco R. Bouckaert, "Naive Bayes Classifiers That Perform Well with Continuous Variables", Advances in Artificial Intelligence, Volume 2871, 2003, pp 326-333.

[5] Han-joon Kim, Jae-young Chang, "Improving Naïve Bayes Text Classifier with Modified EM Algorithm",Advances in Intelligent Data Analysis , Vol 2810, 2003, pp 143-154.

[6] StijnViaene, Richard A. Derrig, and Guido Dedene, "A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis" ,Actions On Knowledge and Data Engineering, Vol. 16, No. 5, May 2004, 612-620.

[7] BojanMihaljevic,Pedro Larrañaga, Concha Bielza, "Augmented Semi-naive Bayes Classifier" ,IEEE Transactions on Systems, Man and Cypernetics-PartB: Cybernetics, Vol. 36, No. 5, Oct 2006, 1149-116.

[8] V. Robles,P. Larrañaga, J. M. Pria, E. Menasalvas, M. S. Perez, " Interval Estimation Naïve Bayes",Advanced Data Mining and Applications", Vol 4632, 2007, pp 134-145

[9] Liangxiao Jiang,Dianhong Wang, ZhihuaCai, Xuesong Yan, "Survey of Improving Naive Bayes for Classification",Advances in Artificial Intelligence,Vol 8109, 2013, pp 159-167

[10] R. R. Ade, Dr. P. R. Deshmukh, "Methods for Incremetnal Learning: A Survey", International Journal of Data Mining & Knowledge Management Proc., *IJDKP)*, 03(04), 119 - 125 July 2013.

[11] Seiichi Ozawa, Shaoning Pang and Nikola Kasabov, Incremental Learning of Chunk Data for Online Pattern Classification Systems, IEEE Transactions on Neural Networks, 2008,1045-9227.

[12] Sheng Uei Guan and Fangming Zhu, An Incremental Approach to Genetic Algorithms Based Classification, IEEE Transactions on Systems Man. And Cybernetics, Vol. 35, No. 2, 2005, 1083-4419.

[13] J. R. Millan, "Rapid, safe, and incremental learning of navigation strategies," IEEE Trans. Syst., Man, Cybern., Part B: Cybern., vol. 26, no. 3, pp. 408–420, Jun. 1996.

[14] G. Y. Chen and W. H. Tsai, "An incremental-learning-by-navigation approach to vision-based autonomous land vehicle guidance in indoor environments using vertical line information and multiweighted generalized Hough transform technique," IEEE Trans. Syst., Man, Cybern, Part B: Cybern., Vol. 28, no. 5, pp. 740–748, Oct. 1998.

[15] Sheng Wan, Larry E. Banta, Parameter Incremental Learning Algorithm for Neural Networks, IEEE Transactions on Neural Networks, Vol. 17, No. 6, 2006, 1045-9227.

[16] Ryan Elwell, RobiPolikar, Incremental Learning of Concept Drift in Non-stationary Environments, IEEE Trans. On NN, Vol 22, No. 10, Oct. 2011.

[17] David Martinez-Rego, Beatriz Perez-Sanchez, Oscar Fontenla-Romero, Amparo Alonso-Betanzos, A robust incremental learning method for non-stationary environments, Neurocomputing 74, 2011.

[18] Sciichi Ozawa, Soon Lee Toh, and Shigeo Abe, Incremental Learning for Online Face Recognition", Proceedings of International Joint Conference on NN, Montreal, Canada, July 31-August 4,2005.

[19] HaitaoZhaYuo and Pong Chi Yuen, Incremental Linear Discriminant Analysis for Face Recognition", IEEE Trans. On Systems, MAN and Cyber., vol. 38, No. 1, Feb,.2008.