Clustering students' based on previous academic performance

Kartik N. Shah[1], Srinivasulu Kothuru[2] and S. Vairamuthu[3]
School of Computing Science and Engineering, VIT University

Abstract

Educational data mining is very popular research area for studying the behavior of students based upon their past performance. As Education is very basic need, which must be given to all, the study of student behavior plays a vital role. Grouping students on the basis of their performance, we can make a good team for any competitions to represent from institute or university. Also we can make some more focus on the students who are having bad performance by giving some extra lectures and motivating them for better study. In this paper, we will cluster the similar behavior students based upon their past academic performance. We are using Similarity measure as Canberra Distance for clustering same type of students. We will use VIT university data for analyzing performance.

Keywords: Educational Data Mining, Clustering Students, Grouping Method, Students behavior.

1. Introduction

Educational Data Mining (EDM) is the study of raw data of educational institutes and universities to get the needed information that can be used by educational software developers, students, teachers, other educational researchers and parents. Moreover, we can say that machine-learning, statistical and data-mining (DM) algorithms are part of EDM to analyze the various types of educational data [3]. EDM is concerned with understanding the student behavior and then create the methods for understanding the other students behavior for creating techniques or algorithms which can help for modifying existing system by which students can understand well [2]. Web Education or E-Learning is the era in which we can use Internet to give education to students. Web Education has created new era of learning and understanding learning pattern of students [4]. This information is gold mine for educational mining [7]. Prediction of student's performance is very useful part of Educational Data Mining. We can use this predicted data in many different contexts in universities. We can identify exceptional students for scholarships or stipend. It is essential part for admission of Post Graduate students. Also we can find the students which can behave poorly in the exams. These predictions will make the management of universities to take remedial measures at early stage which enables institute or university to produce excellent students. Now a day, many universities follow web based education. In this case, it is useful to classify a student's performance. We can use Data Mining (DM) techniques to achieve these objectives [6].

The supervised classification algorithm needs a training sample for each class, which contains the information of known collection of data points of class of interest. So, classification process is based on closer the data point is compared to the training sample. The training samples are representative of the known classes of interest to the analyst. The classification methods which relay on use of training samples are known as supervised classification methods. These methods are basically used for image classification.

The 3 steps involved in supervised classification are:

1. Data training stage: The analyst find the interested area which is already known to him and this known information is used to classify the other unknown data based upon this training data. Best match with training sample is consider in the case of multiple match.
2. The Classification stage: This is the main stage in which, classification is carried out based upon available training data. All the unknown classes are labeled based upon the training data available for classification.
3. Final Output Stage: The final stage is output stage which is concerned with using of data available from classification stage in various applications.

2. Related Work

From the past, we had measure the students' performance by the number of courses they have registered and Grand Point Averages (GPA) or Credit Grand Point Averages (CGPA) they have secured. But, that traditional method that we had used in past, was not guaranteeing that the students have achieved the qualifications required for the next era. Now a day, a student's performance has been evaluated on the basis of the marks given by faculty during written examination. Most of the universities follow the mark based approach in which relative grading is given to all students (marks between 1 to 100) and then categorized in the grades (A, B, C, D, E, F or N) and then nominal scores (1, 2, 3. . .10). And finally linguistic terms like 'Fail' or 'Pass', etc. In this study, a weighted sum of assessment style (i.e each exam have different weight) is used to compute the numerical score of each student based upon the university rules: Quiz (Q), Common Assessment Test (CAT) or Internal (I), Final (F), Performance Appraisals (P). The total of all of the above measures is treated as Final result of the student.

For Intelligent Tutoring System, we can use constraint relaxations and sequential pattern mining to automatically acquire the knowledge [8]. We can predict the student's grade based upon similarity measures called Sum of Absolute Difference and Sum of Squared Difference and then we can assign grades for the system of university which follow the Relative grading scheme [1]. To label the students' behavior, we can use text replays.It is more precise and faster [9]. There can be many different objectives for classifying students based on their characteristics which are 2 steps classification process. In which, first step is to study single classifier accuracy on data and then choose the best and gather this information with weak classifier for voting method [11], In Statistical classification, individual result is taken and grouped into similar results and then majority is considered for classification [12]. Some of the statistical algorithms are linear discriminant analysis, kernel and k nearest neighbors [5]. The decision tree is a set of conditions which is organized in a hierarchical structure [13]. It is a predictive model in which to classify an instance, we need to follow the path in the tree till leaf node. If we are able to reach leaf node then we can classify it on that class. We can convert a decision tree to a set of classification rules. Most well-known decision tree algorithms are C4.5 [13]. Rule Induction - it is an area of machine learning in which, from a set of observations, IF-THEN production rules are extracted [14]. In rule induction, operators correspond to generalization and specialization operations and state corresponds to a candidate rule that transform one candidate rule into another. Examples of rule induction algorithms are CN2 [15], Supervised Inductive Algorithm (SIA) [16], a genetic algorithm using real-valued genes (Corcoran) [17] and a Grammar-based genetic programming algorithm (GGP) [18]. We can use Neural Networks also for rule induction. Examples of neural network algorithms are multilayer perceptron (with conjugate gradient-based training) [19], a radial basis function neural network (RBFN) [20], incremental RBFN [24], decremental RBFN [20], a hybrid Genetic Algorithm Neural Network (GANN) [21] and Neural Network Evolutionary Programming (NNEP) [22].

3. Proposed Framework

Different measures of similarity or distance are convenient for many types of analysis. There are many techniques available to cluster the similar type of students. All grading techniques and classification techniques can be extended for clustering of students. When we classify the students based upon any algorithm, we can conclude that the students who are coming in same class having some type of similar behavior or characteristics. So, we can make a group of students who are coming in same class and make a single cluster of that group based upon the criteria which is used for classification purpose. We will cluster students' records based upon their academic performance of the previous data. Some of the distance and similarity measures for numerical data are listed below. We will use this distance measure to cluster the students based upon their academic performance.

Distance measure between KNOWN(p, q, r) and UNKNOWN(x_j, y_j, z_j)

Euclidean Distance	$\sqrt{\sum_j [Abs(p-x_j)^2 + Abs(q-y_j)^2 + Abs(r-z_j$
Manhattan Distance	$\sum_j [Abs(p-x_j) + Abs(q-y_j) + Abs(r-z_j$
Squared Euclidean Distance	$\sum_j [Abs(p-x_j)^2 + Abs(q-y_j)^2 + Abs(r-z_j$
Bray Curtis Distance	$\sum_j \frac{[Abs(p-x_j) + Abs(q-y_j) + Abs(r-z}{[Abs(p+x_j) + Abs(q+y_j) + Abs(r+z}$
Canberra Distance	$\sum_j \frac{Abs(p-x_j)}{p+x_j} + \frac{Abs(q-y_j)}{q+y_j} + \frac{Abs(r-}{r+z_j}$
Chessboard Distance	$MAX[Abs(p-x_j), Abs(q-y_j), Abs(r-z$
Cosine Distance	$1-\frac{px_j+qy_j+rz_j}{\sqrt{p^2+q^2+r^2}\sqrt{x_j^2+y_j^2+z_j^2}}$

Table1 Similarity measures for numerical data

Table1 shows similarity and distance measures are useful for numerical data. There are many more measures available for finding similarity like Normalized Squared Euclidean Distance, Correlation Distance etc. We can use 1 of the techniques and then apply classification process to create the cluster of similar students. Now days, every university follow their own evaluation patterns based upon that we can modify our approach to calculate the similarity. We will take Canberra Distance as similarity measure and cluster the students based upon their marks obtained in CAT I, CAT II and Quiz. We will study this analysis for VIT University's students' records. In case, if we want to apply the same technique for different university then we can add details in implementation of this technique. We can extent it for analysis of students of final year based upon the previous years. Consider the following details of the students.

CAT1	CAT2	QUIZ	REG_NO
34	36	8	1
14.3	29	9	2
44.5	47	13	3
40	45	14	4
41	46	13	5
45.5	46	15	6
37.5	43	12	7
28.5	43	10	8
38.5	45	13	9
28.5	40	6	10
40	47	14	11
35	40	9	12
19	30	8	13
43.5	46.5	13	14
42.5	44.5	13	15
36	43	12	16
40	40	14	17
41.5	45	14	18
31.5	40	8	19
42	47	13	20
36	38	13	21
42.5	46	14	22
42	33.5	9	23
30	35	13	24
29	30.5	9	25
20	25	7	26
35	41	14	27
38	40	13	28
41	47	15	29
40.5	46	14	30
47.5	43	15	31
25	40	6	32
0	0	0	33
42	41	12	34
28	29	8	35

Table2 Students' Records

After applying above Canberra Distance algorithm, we can cluster the Table2 data. After clustering, we will able to know that which students come in which category. As every universities follow different evaluation patterns, for VIT University, CAT I and CAT II is having 15% weight for each. They conduct 3 quiz and each having 5% of weight. In above records, we have added all 3 quiz marks for simplicity. Then we will calculate the marks based upon the weight of the exam and then all details are together applied to algorithm. We will give reference detail to cluster the data based upon which students will be grouped. We have used following groups to classify students.

Group 'A'	CAT I- 45, CAT II- 45 and Quiz – 15
Group 'B'	CAT I- 40, CAT II- 40 and Quiz – 13
Group 'C'	CAT I- 34, CAT II- 34 and Quiz - 9
Group 'D'	CAT I- 28, CAT II- 28 and Quiz - 8
Unusual Behavior or Failure	Remaining

Table3 Reference marks for clustering

After applying the algorithm, we can get following clusters. Table4 shows the details about clusters after applying algorithms.

Group 'A'	3,4,6,11,14,15,18,22,29,30,31
Group 'B'	4,5,7,9,11,14,15,17,18,20,21,27,28,30,34
Group 'C'	1,12,23
Group 'D'	25,35
Unusual Behavior or Failure	2,8,10,13,16,19,24,26,28,32,33

Table4 Output Clusters

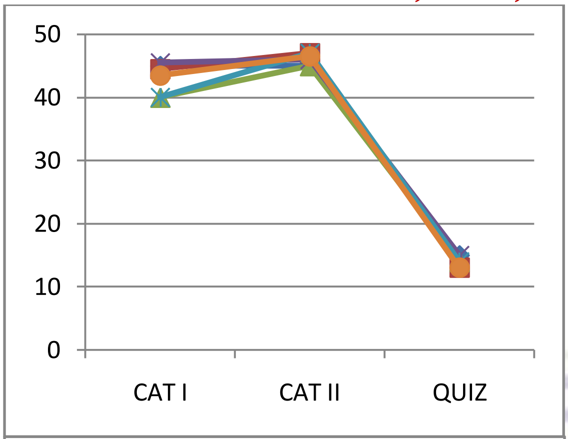
4. Implementation

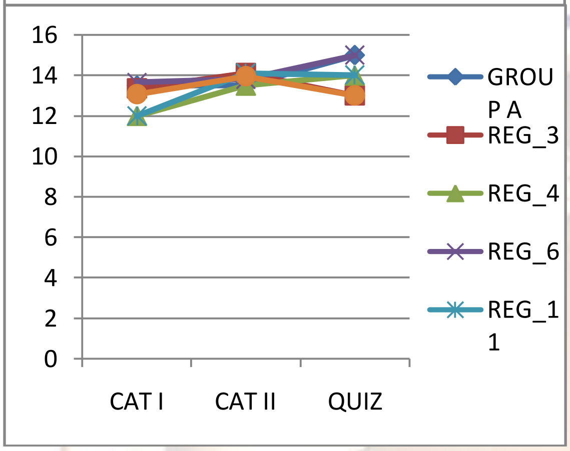
We can implement this algorithm using .Net technology and C# language. It will take all students' records and Reference marks (criteria for clustering) as input and will provide the Reg. no of students who are coming in that cluster. Following figure1 shows the implementation details.

Figure1 Implementation

5. Analysis

After applying algorithm, we can cluster students based upon criteria shown above. We can map in line chart as follows.





For more precision, we can go for normalization to get similar type of points. The line graph can be shown as below. It shows that while we consider the individual weight of the exams then also it can be able to cluster proper way. From Line chart it is clear that all the students are coming in same clusters. So, Canberra Distance algorithm can group the students based upon similar academic performance in previous exams.

6. Future Work

In this paper, we have discussed marks as a measure for clustering the students. There are many other behaviors by which we can cluster the students. Some of the measures are based upon practical knowledge, class behavior, talent in particular field, family background. We have not considered any of the measures as it is very difficult to understand each student. We can enhance the clustering techniques based upon the above mentioned criteria.

7. Conclusion

From the above analysis, we can say that we can make cluster of students based upon Canberra Distance Similarity and Distance measure. In this study, we have use VIT University data as a reference to check validity of results. We can extend this algorithm for any university which follow numerical evaluation students (i.e based upon marks). Most of the universities follow the marks pattern and then marks are converted into grades. So, this analysis can be applied on most of the university's students' records.

8. Acknowledgement

The authors would like to thank the School of Computer Science and Engineering, VIT University, for giving them the opportunity to carry out this project and also for providing them with the requisite resources and infrastructure for carrying out the research.

References

[1] Kartik N. Shah, Shantanu Santoki, Himanshu Ghetia, L.Ramanthan, "Mining on Student's Records to Predict the Behavior of Students", IEEE - International Conference on Research and Development Prospects on Engineering and Technology, March 29,30 - 2013 Vol.4, pp. 54-57.

[2] R. Baker, "Data mining for education," in *International Encyclopedia of Education*, B.McGaw, P. Peterson, and E. Baker, Eds., 3rd ed. Oxford, U.K.: Elsevier, 2010.

[3] T. Barnes, M. Desmarais, C. Romero, and S. Ventura, presented at the 2nd Int. Conf. Educ. Data Mining, Cordoba, Spain, 2009.

[4] F. Castro, A. Vellido, A. Nebot, and F. Mugica, "Applying data mining techniques to e-learning problems," in *Evolution of Teaching and Learning Paradigms in Intelligent Environment* (Studies in Computational Intelligence), vol. 62, L. C. Jain, R. Tedman, and D. Tedman, Eds. New York: Springer-Verlag, 2007, pp. 183–221.

[5] Minaei-Bidgoli, B., Punch, W. Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System. Genetic and Evolutionary Computation, Part II. 2003. pp.2252–2263.

[6] Romero, C., Ventura, S. Educational Data Mining: a Survey from 1995 to 2005. Expert Systems with Applications, 2007, 33(1), pp.135-146.

[7] J. Mostow and J. Beck, "Some useful tactics to modify, map and mine data from intelligent tutors," *J. Nat. Lang. Eng.*, vol. 12, no. 2, pp. 195– 208, 2006.

[8] C. Antunes, Acquiring background knowledge for intelligent tutoring systems, in: Proceedings of the 2nd International Conference on Educational Data Mining, 2008, pp. 18–27.

[9] Baker, R.S.J.D. and De Carvalho, A.M.J.A. 2008. Labeling Student Behavior Faster and More Precisely with Text Replays. In *Proceedings of the 1st International Conference on Educational Data Mining*, 38-47

[10] Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy "A Comparative Analysis of Techniques for Predicting Academic Performance" 37th ASEE/IEEE Frontiers in Education Conference, October 10 – 13, 2007, Milwaukee, WI, T2G-7

[11] Muhammad Sufyian Bin Mohd Azmi "Academic Performance Prediction Based On Voting Technique" IEEE 2011.

[12] Otero, J., Sánchez, L. Induction of Descriptive Fuzzy Classifiers with the Logitboost Algorithm. Soft Computing 2005, 10(9), pp.825-835.

[13] Quinlan, J.R. C4.5: Programs for Machine Learning. Morgan Kaufman. 1993.

[14] Delgado, M., Gibaja, E., Pegalajar, M.C., Pérez, O. Predicting Students' Marks from Moodle Logs using Neural Network Models. Current Developments in Technology- Assisted Education, Badajoz, 2006. pp.586-590.

[15] Clark, P., Niblett, T. The CN2 Induction Algorithm. Machine Learning 1989, 3(4), pp.261-283

[16] Venturini, G. SIA A Supervised Inductive Algorithm with Genetic Search for Learning Attributes based Concepts. Conf. on Machine Learning, 1993. pp.280-296.

[17] Corcoran, A.L., Sen, S. Using Real-valued Genetic Algorithms to Evolve Rule Sets for Classification. Conference on Evolutionary Computation, Orlando, 1994, pp.120-124.

[18] Espejo, P.G., Romero, C., Ventura, S. Hervas, C. Induction of Classification Rules with Grammar-Based Genetic Programming. Conference on Machine Intelligence, 2005. pp.596-601.

[19] Moller, M.F. A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. Neural Networks, 1993, 6(4), pp.525-533.

[20] Broomhead, D.S., Lowe, D. Multivariable Functional Interpolation and Adaptative Networks. Complex Systems 11, 1988, pp.321-355.

[21] Yao, X. Evolving Artificial Neural Networks. Proceedings of the IEEE, 1999, 87(9), pp.1423-1447.

[22] Martínez, F.J., Hervás, C., Gutiérrez, P.A., Martínez, A.C., Ventura, S. Evolutionary Product-Unit Neural Networks for Classification. Conference on Intelligent Data Engineering and Automated Learning. 2006. pp.1320-1328.