International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Academic Performance Evaluation and Prediction using SC-FCM Clustering Algorithm

Govinda Raj S S

Student, B.Tech Information Technology, Madras Institute of Technology, Anna University, Chennai, Tamil Nadu, India

Abstract: ***In this Paper I've introduced a methodology to evaluate and predict the academic performance of Students using Subtractive Clustering – Fuzzy C means Clustering Hybrid model taking their previous performance as training data. Students' improvement in performance is henceforth evaluated using the Algorithm predicted results and original results.***

Keywords: Academic Performance Evaluation, Subtractive Clustering, Fuzzy C means clustering, training data

1. Introduction

Evaluation of Students' academic performance can be considered as a clustering problem where clusters are formed based on intelligence of students (the grades they secure in previous results). In most of the Higher Education Institutions, a Student is examined by Assessment Tests, Assignments, and Projects and finally given an overall grade for the Credited Paper in that Semester. Hence, the academic performance of students can be determined by the final grade points in each semester.

2. Data Clustering Algorithm for Academic Performance Prediction

Data Clustering Algorithms allocate students into homogeneous groups depending upon their performance in previous semesters, starting from minimum observance period (Say first two semesters).

Fuzzy C-means Clustering

This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. Iteratively, membership and cluster centers are updated according to the formula:

$$J(U,V)=\sum_{i=1}^{k}\sum_{x_k\in X}(\mu_{C_i}(x_k))^m \,\|x_k-v_i\|^2, \qquad (1)$$

The membership matrix U is allowed to have elements with values between 0 and 1. However, the summation of degrees of belongingness of a data point to all clusters is always equal to unity.

$$\mu_{C_i}(x)=\frac{1}{\sum_{j=1}^{k}\left(\frac{\|x-v_i\|^2}{\|x-v_j\|^2}\right)^{\frac{1}{m-1}}} \qquad 1\le i\le k,\ x\in X, \qquad (2)$$

$$v_i=\frac{\sum_{x\in X}(\mu_{C_i}(x))^m x}{\sum_{x\in X}^{n}(\mu_{C_i}(x))^m} \qquad 1\le i\le k, \qquad (3)$$

$$\sum_{i=1}^{C}\|v_i^{\text{Previous}}-v_i\|\le\varepsilon. \qquad (4)$$

Subtractive Clustering

Number of cluster centers and their initial locations are pre-specified in algorithms like k-means clustering and Fuzzy C-means clustering. The quality of the solution depends strongly on the choice of initial values. Yager and Filev [7] proposed a simple and effective algorithm, called the mountain method, for estimating the number and initial location of cluster centers. Their method is based on gridding the data space and computing a potential value for each grid point based on its distances to the actual data points. The grid point with the highest potential value is chosen as the first cluster center. The key idea in their method is that once the first cluster center is chosen, the potential of all grid points is reduced according to their distance from the cluster center. Although this method is simple and effective, the computation grows exponentially with the dimension of the problem because the mountain function has to be evaluated at each grid point.

Chiu [8] proposed an extension of mountain method, called subtractive clustering. This method solves the computational problem associated with mountain method. It uses data points as the candidates for cluster centers, instead of grid points as in mountain clustering. The computation for this technique is now proportional to the problem size instead of the problem dimension.

Subtractive clustering treats each point as a potential cluster centre and uses the following equation

$$D_i = \sum_{j=1}^{n} \exp\left[\frac{|x_i - x_j|^2}{\left(\frac{r_a}{2}\right)^2}\right], \quad (5)$$

where r_a defines the neighborhood radius for each cluster is Euclidean distance and n is the number of sampling points of the dataset x. Using eq. (5), subtractive algorithm computes the potential for each point. The point with the highest potential, denoted by D_{c1} is selected as the first cluster centre x_{c1}. Next, the potential of each data point x_i is updated as follows

$$D_i = D_i - D_{c1} \exp\left[\frac{|x_i - x_{c1}|^2}{\left(\frac{r_b}{2}\right)^2}\right], \quad (6)$$

where r_b represents the radius of the neighborhood with significant potential reduction. This process continues till a stopping criterion is met.

Hybrid SC-FCM Algorithm

The hybrid SC-FCM algorithm is presented as follow:

1) Calculate the density of every data point using eq. (5), and the highest density of the point is chosen as x_{c1}.
2) Set $n_c = 1$, consider the highest potential of data point as D_{c1} with location as x_{c1} for the first cluster centre.
3) Update each point potential using eq. (6).
4) The results of aforementioned including the clustering number and cluster centers are chosen as the FCM initial values. The initial fuzzy partition matrix U(0) is also set contemporaneously as follows:

$$\mu_{ik}(0) = \frac{1}{\sum_{j=1}^{c} \left(\frac{D_{ik}}{D_{jk}}\right)^{\frac{2}{m-1}}} \quad i=1,2,\ldots,c \quad k=1,2,\ldots N \quad (7)$$

where D_{jk} , which is calculated firstly, signifies the distances between k^{th} data point and j^{th} initial cluster center.

5) Calculate the center values according to eq.(3)
6) Update the fuzzy partition matrix U(k) according to eq.(2)
7) If eq.(4) is satisfied, then stop; otherwise, k = k + 1, return to step (5)

3. Formulation

3.1 Generation of Training Data

Students' academic performance is more or less the same when compared to the previous batches. Results of Students in the past batches are collected / generated to divide them into groups. The Academic grade point result of each student in the previous batches is given as training data. The model is developed taking his/her grade points obtained in each subjects as features. Thus training set is further divided into clusters using SC-FCM Clustering Algorithm discussed below.

3.2 Applying SC-FCM Clustering Algorithm to the Generated Data

The generated data is then passed to the clustering algorithm. Threshold for the algorithm to stop is set proportional to the size of training data.

Optimal Grade of Students in each Cluster is obtained for every Subject. A matrix representing the grade of students corresponding to the subjects is formed for each cluster.

Most of the higher education institutions represents grade using a grade point value. For calculation purpose grade point is chosen here instead of the abstract grade. The grade point secured by all the students trained in a same cluster are averaged for each subject and updated in the matrix (Subject – Grade Matrix)

The precision of the floating point number representing the grade point value is minimized to two.

3.3 Testing Phase

The Algorithm developed with the training data is tested for accuracy and deviation is measured. This ensures the correctness of the Algorithm. Moreover Over fitting or under fitting can be easily identified here.

With the grade points obtained by the students for a specified span of observance period (say first and second semesters), the Algorithm groups the student into the nearest suitable cluster. While grouping the students, a mediocre grade point value is allocated to each subject in the other semesters, whose values are not yet known. This is required because clustering can be done only when all the fields in the feature set is available for every record in the training data.

Then the results of the immediate next semester (say 3^{rd} semester incase first two semester results are known) is predicted. Here prediction is done by clustering the student record with already clustered records. Nearest cluster is identified and the subject grade points on these semesters are taken as predicted results.

The difference in the results (original - predicted) is measured as prediction error. The error is measured as a floating point number between 0 and 1.

The algorithm is again trained to reduce this prediction error.

4. Evaluation and Prediction

With the grade points obtained by the students for a specified span of observance period (say first and second semesters), the Algorithm groups the student into the nearest suitable cluster. Similar to the case of testing phase, the variance in the grade point of a student and that allocated to the clustered group is determined.

Then the results of the immediate next semester (say 3^{rd} semester incase first two semester results are known) is predicted. Here prediction is done by clustering the student record with already clustered records. Nearest cluster is identified and the subject grade points on these semesters are taken as predicted results.

With the common general grade point predicted for the whole set, variance determined for individual student is added.

After the original results being declared, the performance increase / decrease / neutrality is identified using the change in the results.

5. Algorithm to predict the Academic Performance of Students

1) The training data (results of students from previous batch) are clustered using SC-FCM clustering Algorithm.
2) Optimal grade point value for each subject is calculated for each cluster by averaging the grade point of each student obtained in the subject. This is maintained in cluster subject matrix.
3) The test set of results are then included with only the results of observance period given. For the remaining subjects a mediocre grade point value substituted for the cluster is substitutes and clustered.
4) Then the results of the immediate next semester are predicted from the cluster's substituted grade point value.
5) From the actual result and the predicted result, prediction error is calculated. If the error is very high, the algorithm is trained with a different set of training data.
6) In the evaluation and prediction phase, the same process handled by testing is repeated but with the students whose performance needs to be predicted.
7) During the observance period, the students' grade may vary a little from the cluster's average result. The difference is calculated as variance.
8) Then the results of the immediate next semester are predicted from the cluster's substituted grade point value and the variance calculated for each student.
9) From the actual result and the predicted result, the students' performance increase / decrease / neutrality is predicted.

Experimental Discussions

In the following tables, students' academic performance is evaluated in terms of grade point (Scale: 5 – 10)

Table 1: Describing Training Data with a set of 40 students and their grade point in 5 subjects

Subject	*A*	*B*	*C*	*D*	*E*
Student 1	10	9	8	9	8
Student 2	8	8	7	7	8
Student 3	6	6	5	6	6
Student 4	5	7	10	10	9
Student 5	10	10	8	9	9
Student 6	7	7	8	7	8
Student 7	7	7	5	5	6
Student 8	6	6	9	10	10
Student 9	9	9	8	9	9
Student 10	8	7	6	8	9
Student 11	7	6	5	7	6
Student 12	7	6	9	9	9
Student 13	10	9	9	8	8
Student 14	9	8	7	7	8
Student 15	6	6	5	7	6
Student 16	5	6	8	9	10
Student 17	9	10	10	9	9
Student 18	9	8	7	7	8
Student 19	5	6	6	7	6
Student 20	6	7	8	10	10
Student 21	8	9	9	8	9
Student 22	8	7	6	8	9
Student 23	7	6	6	5	5
Student 24	7	5	8	10	8
Student 25	10	9	9	9	9
Student 26	9	8	7	8	9
Student 27	6	6	6	7	6
Student 28	6	8	9	9	9
Student 29	9	10	8	10	8
Student 30	9	7	7	7	8
Student 31	7	5	6	5	7
Student 32	5	7	10	10	10
Student 33	10	10	8	10	8
Student 34	8	8	6	8	9
Student 35	5	6	7	6	5
Student 36	6	5	9	8	9
Student 37	8	9	9	8	8
Student 38	8	7	6	8	9
Student 39	6	7	5	6	7
Student 40	6	7	10	10	9

Cluster – Subject Matrix obtained by running the clustering algorithm is consolidated as follows:

Table 2: Describing Cluster-Subject Matrix after processing the clustering algorithm to Table 1

Subject	A	B	C	D	E
Cluster 1	9.3	9.4	8.6	8.9	8.5
Cluster 2	8.3	7.5	6.7	7.5	8.5
Cluster 3	6.2	6.1	5.6	6.1	6
Cluster 4	5.9	6.4	9	9.5	9.3

Table 3: Describing Testing Data with a set of 10 students and their grade point in 5 subjects

Subjects	A	B	C	D	E
Student 1	9	8	7	8	9
Student 2	6	6	6	7	6
Student 3	6	8	9	9	9
Student 4	7	5	6	5	7
Student 5	5	7	10	10	10
Student 6	10	10	8	10	8
Student 7	8	8	6	8	9
Student 8	6	5	9	8	9
Student 9	8	9	9	8	8
Student 10	8	7	6	8	9

Test Set is trained with only two subject's grade points A and D. The remaining subjects are given the average value of these two subjects. Clustering of the Testing Data is done in the following way:

Table 4: Cluster formation of the Test Set

Cluster 1	Students 6 and 9
Cluster 2	Students 1,7 and 10
Cluster 3	Students 2 and 4
Cluster 4	Students 3,5 and 8

Table 5: Original Grade Points Obtained by Students whose results are to be predicted

Subjects	A	B	C	D	E
Student 1	10	10	8	9	9
Student 2	7	7	8	7	8
Student 3	6	6	9	10	10
Student 4	8	7	6	8	9
Student 5	7	6	5	7	6
Student 6	7	6	9	9	9
Student 7	10	9	9	8	8
Student 8	6	6	5	7	6
Student 9	5	6	8	9	10
Student 10	9	10	10	9	9
Student 11	9	8	7	7	8
Student 12	5	6	6	7	6
Student 13	8	9	9	8	9
Student 14	7	6	6	5	5
Student 15	7	5	8	10	8

Table 6: Predicted Results

Subjects	A	B	C	D	E
Student 1	**10**	9	8	**9**	9
Student 2	**7**	7	8	**7**	8
Student 3	**6**	6	9	**10**	9
Student 4	**8**	6	6	**8**	9
Student 5	**7**	6	5	**7**	6
Student 6	**7**	6	9	**9**	9
Student 7	**10**	9	9	**8**	8
Student 8	**6**	7	5	**7**	7
Student 9	**5**	6	8	**9**	10
Student 10	**9**	10	9	**9**	9
Student 11	**9**	8	7	**7**	8
Student 12	**5**	5	6	**7**	6
Student 13	**8**	9	9	**8**	9
Student 14	**7**	6	6	**5**	6
Student 15	**7**	6	8	**10**	8

Average deviation in grade point prediction is about 0.33 in Subject B, 0.067 in Subject C, 0.2 in Subject E. Their Performance is evaluated increase or decrease in the following table:

Table 7: Performance Evaluation based on Predicted and Original Results

Subjects	B	C	E
Student 1	I	N	N
Student 2	N	N	N
Student 3	N	N	I
Student 4	I	N	N
Student 5	N	N	N
Student 6	N	N	N
Student 7	N	N	N
Student 8	D	N	D
Student 9	N	N	N
Student 10	N	I	N
Student 11	N	N	N
Student 12	I	N	N
Student 13	N	N	N
Student 14	N	N	D
Student 15	D	N	N

I – Increase
N – Neutrality
D – Decrease

6. Conclusion

Hence the results of the upcoming semesters for a set of students can be predicted and their performance can be evaluated. This algorithm purely relies on the correctness of the training data.

Predicting the performance of students helps them to be aware of their development in terms of academic activities.

References

[1] Huaguang, Z., Jianhong, L. and Laijiu, C., The theory and application research on predictive generalized control. Acta Autom. Sin., 1993, 19(3), 9–17
[2] Oyelade, O. J., Oladipupo, O. O. and Obagbua, I. C., Application of K-means clustering algorithm for prediction of students' academic performance. Int. J. Comput. Sci. Inf. Security, 2010, 7(1), 292–295.
[3] Ramjeet Singh Yadav, P.Ahmed, A.K.Soni and Saurabh Pal,"Academic performance evaluation using soft computing techniques",CURRENT SCIENCE, VOL.106,NO.11,10 JUNE 2014, 1505-1517
[4] Sansgiry, S. S., Bhosle, M. and Sail, K., Factors that affect academic performance among pharmacy students. Am. J. Pharm. Educ., 2006, 70(5), 231–243.
[5] Sarala.N, Kavitha.R."Evaluation Of Students' Mathematical Skill Using Fuzzy Inference System" Proceeedings of International Conference On Recent Advance IN Applied Science, Feb 11-13,2016,Vol.2, 497-503.
[6] Hossen, J., Rahman, A., Sayeed, S., Samsuddin, K. and Rokhani, F., A modified hybrid fuzzy clustering algorithm for data partitions. Austr. J. Basic Appl. Sci., 2011, 5(8), 674–681.

Author Profile

Govinda Raj S S currently pursuing B.Tech. degree in Information Technology from Madras Institute of Technology, Anna University, Chennai, Tamil Nadu, India