International Journal of Current Advanced Research
ISSN: O: 2319-6475, ISSN: P: 2319-6505, Impact Factor: 6.614
Available Online at www.journalijcar.org
Volume 7; Issue 4(E); April 2018; Page No. 11669-11672
DOI: http://dx.doi.org/10.24327/ijcar.2018.11672.2026

Research Article

GENERATION AND ANALYSIS OF EDUCATIONAL DATASETS USING k-MEANS CLUSTERING

Gauri Shanker Kushwaha* and Bharat Mishra

M.G.C.G.V., Chitrakoot, Satna, M.P.

A R T I C L E I N F O

Article History:

Received 18th January, 2018
Received in revised form 13th February, 2018 Accepted 15th March, 2018
Published online 28th April, 2018

Key words:

k-means clustering, WEKA 3.8.1, Data Generation Methods, Instance of a Cluster.

A B S T R A C T

Educational Data Mining is a growing field exploring data in educational perspective by applying diverse data mining tools. It provides built-in knowledge of teaching and learning method for successful education preparation. In the history of data mining, *k*-means algorithm plays an important role because of its extensive implementation. Data mining in educational datasets can be applied to discover patterns/clusters in the untrusted datasets to computerize the decision making process of institutional actors responsible for improving the higher educational quality. *k*-Means is an algorithm to classify the objects on the basis of features of attributes in K number of groups, where K indicates a positive integer. The proposed study is a combination of data mining algorithm and the new concepts of data mining, which aims to improve the quality by combining the institutional actor's data with the knowledge which has been extracted from databases, and providing the precise advice to the concern in scientific manner.

INTRODUCTION

A large number of educational institutes that adopted an information system has been increasing quickly in recent years, one after the other, the amount of data available in each educational institute's databases were also improved. Educational data mining is spontaneously useful to discover knowledge in order from this data that would develop the quality of the whole educational system. Educational data mining can be applied to discover patterns in un–trusted datasets to computerize the decision making process of students, teachers, institutional leaders and quality teaching units in terms of institutional actors of higher education. Educational data which indicates the learning patterns of students in different forms and with various accuracy levels. Based on these results the teachers can provide the necessary guidance to the students who needed more attention and also as an assistance to improve their capabilities on teaching and this will enable the knowledge producers to dynamically change the knowledge flows within the e-learning environments in a more effective and efficient manner. A recently introduced concept of academic analytics uses the data mining algorithms for the educational data of students and gives certain insights about the expected performances of the students, expected retention rate of students and percentage of resources properly utilized.

Corresponding author:* **Gauri Shanker Kushwaha
M.G.C.G.V., Chitrakoot, Satna, M.P.

Data mining is the process of taking information from a data set and convert it into an understandable and meaningful structure for further use. There are various techniques of data mining like classification, clustering, association rule mining etc. each technique has its own importance according to his role. In this paper clustering technique has been used for further study.[1]

k– Means clustering is an algorithm to classify the objects based on attributes/features into K number of group where K is positive integer number. The basic step of *k*-means clustering is simple and easier to use. [1] Minimizing sum of squares of distances between data and the corresponding cluster centroid grouping is done and the intention is to classify the data. These results also facilitate administrators in decision making and answering certain questions like whether the faculty v/s student's ratio is giving satisfactory results or there is a change needed in the teaching methodology. In this study data mining algorithms were applied on educational datasets and analyzed for exploring the hidden knowledge from generated dataset.

METHODOLOGY

For identification of quality of higher education by applying data mining algorithm on generated data set and output of this analysis is used for improvement in available condition of quality of higher education. The flowchart shown in figure 1 defines the overall process of units with set of instructions that are:

Step-1: Determining the factors of higher education.
Step-2: Collect the responders of institutional actors.

Step-3: Organize the data for pre- processing in interface.
Step-4: Selection of the algorithm.
Step-5: Preprocess the data for result.
Step-6: Visualize the obtain result.
Step-7: Compare the result obtained with existing model.
Step-8: Get the result.

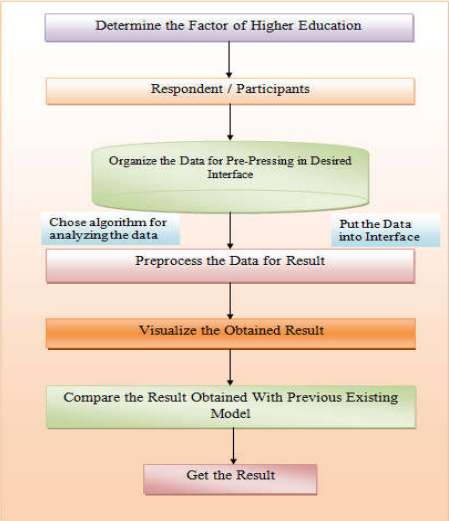

Figure 1 Flowchart of data generation and analysis using k- means clustering to improve higher educational quality

Selected data mining algorithms are considered for analyzing on taken data set. These algorithms are executed on taken data set (primary dataset) by the help of WEKA 3.8.1 software. The data is preprocessed according to this base file which is accepted by WEKA 3.8.1 software. The generated data is converted into comma-separated format (.csv) from excel (.xls) format with file name.csv because WEKA 3.8.1 does not identify the excel data. Further data is preprocessed according WEKA software, after preprocessing the file is converted in file_name.arff of preprocessed data file, the data file of WEKA converted was initialized for the visualization of cluster.

One of the challenges of *k*-means algorithm is finding an optimal initial cluster assignment and an optimal K value (number of cluster centers) to minimize the sum of squared errors. Rerun the algorithm with different seed and numClusters values and compare the error rates. Increasing K reduces the distances to cluster centers, and the sum of squared errors converges. The algorithm was implemented in WEKA software and the output of implemented algorithm is in graphical representation of quality of higher education from taken data set.

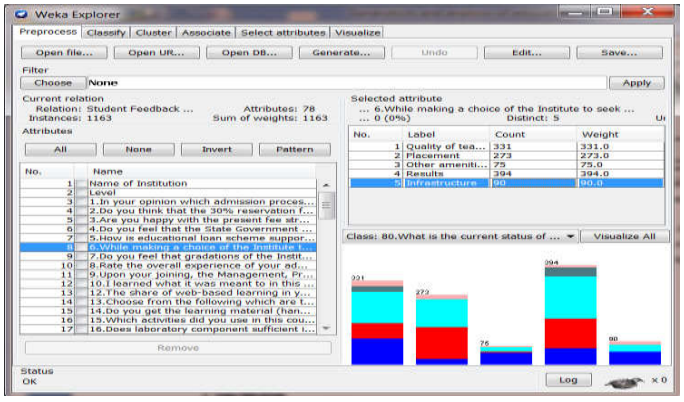

Figure 2 Preprocessing of Higher Educational Dataset

In the graphical representation by different colored instances of output file were identified. Significance of different color attributes is different. From color representation of clusters, it is easily clear which cluster having more instances is more significant in particular cluster. The preprocessing of the file students feedback is shown in above figure 2. The further more files were preprocessed in the same way.

k-Means only allow numerical values for attributes, in that case, it may be necessary to convert the data set into the standard spreadsheet format and convert definite attributes to binary. It may also be necessary to normalize the values of attributes that are measured on substantially different scales. While WEKA provides filters to accomplish all of these preprocessing tasks, they are not necessary for clustering in WEKA. WEKA, Simple *k*-means algorithm automatically handles a mixture of categorical and numerical attributes. Furthermore, the algorithm automatically normalizes numerical attributes when doing distance computations. The WEKA Simple *k*-means algorithm uses Euclidean distance measure to compute distances between instances and clusters.

DATA GENERATION AND ANALYSIS

The fact that the classroom environment is dynamic, individualistic and multivariate requires generation of data that can provide the most complete understanding of the instructional process. Analyzing what goes on in classrooms, therefore, requires a systematic approach including the specific research steps of data collection, data analysis and as well as interpretation and application of findings.

The data was recorded through questionnaire from different academic/research institutions and it also includes an online survey of various respondents of 1163 students, 170 teachers, 46 institutional leaders and 11 quality teaching units, that includes a variety of various features of attributes.

By minimizing sum of squares of distances between data and the corresponding cluster centroid grouping is done and the intention is to classify the data. The basic step of *k*-means clustering is simple and easier to use. In the beginning a number of K cluster determined and assumed that the centroid or center of these clusters. The algorithm can take any random objects as the initial centroid or the first K objects in sequence can also serve as the initial centroid and will do the below given steps until convergence iterate until constant:

1. Determine the centroid coordinate
2. Determine the distance of each object to the centroid
3. Group the object based on minimum distance

As an illustration of performing clustering in WEKA, we will use its implementation of the *k*-means algorithm to cluster the respondents in this educational data set, and to characterize the resulting respondent's segments.

Below given figure shows the main WEKA Explorer interface with the data file loaded. For the analysis, the preprocessing of datasets has been performed which clarifies the relational information that includes number of attributes and instances, whereas after analysis number of iteration and time taken for the analysis, shown in table 1.

Table 1 Relational and methodical information of the datasets

SN	Information	Student feedback	Teacher feedback	Institutional Leaders feedback	Quality Teaching feedback
1	No. of attributes	78	56	69	40
2	No. of instances	1163	170	45	11
3	Taken time(Second)	0.27	0.05	0.03	0.00
4	No. of iteration	07	03	03	02

The three different dimension's available x-axis, y-axis, and color are selected for obtaining the cluster number and any of the other attributes. Different combinations of choices will result in a visual rendering of different relationships within each cluster. In the below given result, we have chosen the instance number as the x-axis, Name of the institute as the y-axis, and the colour (cluster) attribute as the color dimension. This will result in a visualization of the distribution of each instance in each cluster. As an illustration of performing clustering in WEKA, we have used its *k*-means algorithm to create the cluster of the respondents in this educational data set within dataset, and to characterize the resulting respondent's segments. Below given figure 3, 4, 5 and 6 shows each cluster through visualization interface of file name "student_feedback.csv", "teacher_feedback.csv", qualityteaching_feedback.csv" and "institutional_feedback.csv" simultaneously.

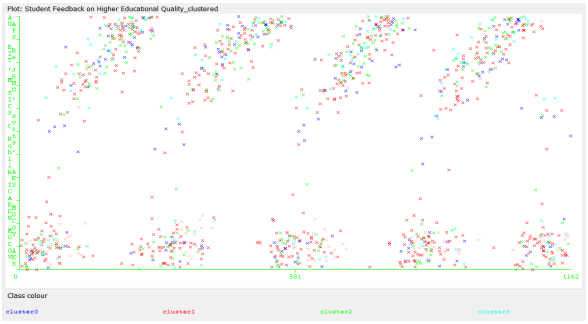

Figure 3 Students feedback

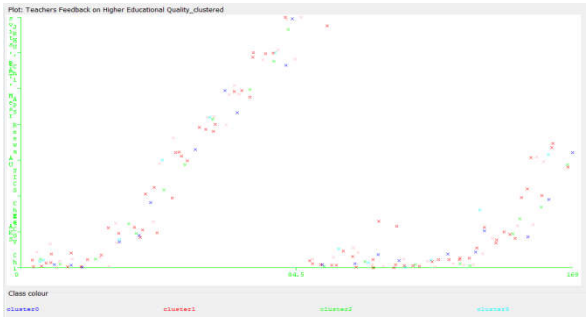

Figure 4 Teachers feedback

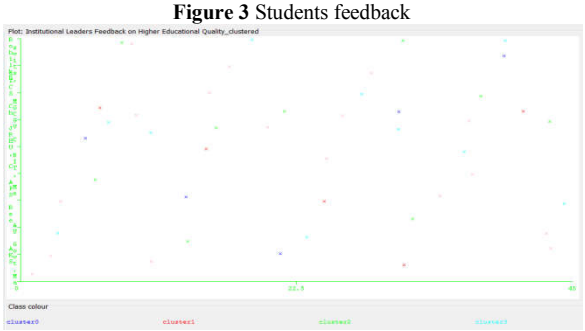

Figure 5 Institutional Leaders feedback

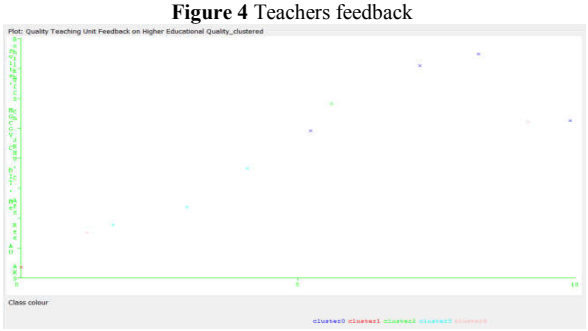

Figure 6 Quality teaching units feedback

Table 2 Total number of instance and percentage of attributes in each dataset

Name of Cluster	Students		Teachers		Institutional leaders		Quality teaching unit	
	No. of instances	Percentage	No. of instances	Percentage	No. of instances	Percentage	No. of instances	Percentage
Cluster_0	154	13	21	12	05	11	04	36
Cluster_1	551	47	74	44	05	11	01	09
Cluster_2	206	18	21	12	09	20	01	09
Cluster_3	154	13	12	07	10	22	03	27
Cluster_4	98	8	42	25	17	37	02	18

The above visualization results the centroid of each instances with their distance of initial object centroid. The result buffer of the build model with the percentage of each cluster and number of attributes categorization was given in table 2, that explains the result itself and depicted graph decides the performance of the quality of higher education in respect of level of students that clarifies the cluster_1 as most preferable by the institutional actors.

From the result it is obtained that in the analysis of educational dataset of file name "student_feedback.csv" and "teacher_feedback.csv" the cluster_1 represents maximum number of instances, "qualityteaching_feedback.csv" cluster_0 and in the "institutional_feedback.csv" cluster_4 represents maximum number of instances.

RESULT & DICSUSSION

In this study the analysis was done using *k*-means clustering algorithm with the generated datasets from the institutional actors of higher education. The maximum responses were recorded in student's feedback dataset which was shown in table 2. The table 2 also represents the maximum number of instances in teachers, institutional leaders and quality teaching unit's feedback as 74, 17 and 04 simultaneously. The overall performance of the analysis on the basis of recorded responses and analysis, Cluster_1 was identified as the maximum clustered instance of the attributes. The below given table 3 represents the total numbers of the recorded instances in each cluster and depicted graph (see figure 7) also clarifies the cluster_1 as the maximum clustered instance.

The result was compared with the pre-existing scale of higher educational quality, NAAC, which is categorized in four major categories of 'very good', 'good', 'satisfactory' and 'below threshold'. In the study the decision parameters were

preconfigured at cluster_0 and which indicates the category first of the scale and further more were in decreasing order of scale and clusters.

Table 3 The total numbers of the recorded instances in each cluster

SN	Institutional actors	Cluster_0	Cluster_1	Cluster_2	Cluster_3	Cluster_4
1	Students	154	551	206	154	98
2	Teachers	21	74	21	12	42
3	Institutional Leaders	05	05	09	10	17
4	Quality Teaching Unit	04	01	01	03	02
	Total	184	631	237	179	159

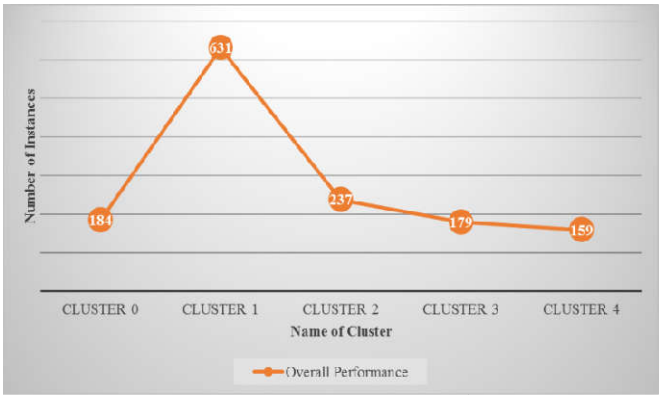

Figure 7 Showing Maximum number of attributes within clusters

CONCLUSION

The study finds out a unique status of higher education that categories the affecting factors as well as institutional actors itself. It also recognizes the maximum number of respondents in the cluster or maximum number of instance in a cluster. The data generated for the analysis was recorded through the survey on the basis of questionnaire. The questionnaire was prepared with the help of focus group and they provided the research experience in own manner to complete the survey. The algorithm categorizes five clusters including different instances. Each includes a number of respondent that subject to an instance.

On the basis of above educational dataset analysis in the perspective of student and teacher feedback it is concluded that quality of higher education of the selected institutions is categorized as good and from quality teaching unit indicates very good whereas in the respect of institutional leaders is stated as below threshold. These categories were derived from the dataset analysis and its overall comparison of these dataset shows a healthy status of the higher education.

Refrences

1. Kushwaha G. S. and Mishra B. (2017), Data Mining Algorithm used in Educational datasets, *International Journal of Scientific Research,* Volume 6, Issue 2, 676-677.
2. Gill P., Stewart K., Treasure E. and Chadwick B. (2008), Methods of data collection in qualitative research: interviews and focus groups, *British Dental Journa*l, Volume 204, No. 6, 291-295.
3. May K. M., (1991), Interview techniques in qualitative research: concerns and challenges. *In Morse J M (ed) Qualitative nursing research, Newbury Park: Sage Publication,* 187-201.
4. Britten N., (1999), Qualitative interviews in healthcare in Pope C, *Mays N (eds) Qualitative research in health care. London: BMJ Books*, Volume 2, pp 11-19,.
5. Morgan D. L., (1998), The focus group guide book. *London: Sage Publications*.
6. Bloor M., Frankland J., Thomas M. and Robson K. (2001), Focus groups in social research. *London: Sage Publications.*
7. Stewart D. W. and Shamdasani P. M., (1990), Focus groups, Theory and practice, *London: Sage Publications.*
8. Dr. Mishra B. and Kushwaha G. S., (2016), A Review of Quality Factors of Higher Education, IOSR *Journal of Research & Method in Education*, Volume 6, Issue 4, 62-68.

How to cite this article:

Gauri Shanker Kushwaha and Bharat Mishra (2018) 'Generation and analysis of educational datasets using k-means clustering', *International Journal of Current Advanced Research,* 07(4), pp. 11669-11672.
DOI: http://dx.doi.org/10.24327/ijcar.2018.11672.2026

****