REVIEW ARTICLE OPEN ACCESS

Study and Analysis of K-Means Clustering Algorithm Using Rapidminer

A CASE STUDY ON STUDENTS' EXAM RESULT

Abhinn Pandey
Computer Technology, Kavikulguru Institute of Technology and Science Nagpur, India

Abstract
Institution is a place where teacher explains and student just understands and learns the lesson. Every student has his own definition for toughness and easiness and there isn't any absolute scale for measuring knowledge but examination score indicate the performance of student. In this case study, knowledge of data mining is combined with educational strategies to improve students' performance. Generally, data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for data. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational database. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).This project describes the use of clustering data mining technique to improve the efficiency of academic performance in the educational institutions .In this project, a live experiment was conducted on students .By conducting an exam on students of computer science major using MOODLE(LMS) and analysing that data generated using RapidMiner(Datamining Software) and later by performing clustering on the data. This method helps to identify the students who need special advising or counselling by the teacher to give high quality of education.

Keywords: Data mining, Clustering, k-means, Moodle, RapidMiner, LMS (Learning Management System)

I. Introduction

Data mining, also called knowledge discovery in databases, in computer science, is the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyze large digital collections, known as data sets. Data mining is widely used in business (insurance, banking, retail), science research (astronomy, medicine), and government security (detection of criminals and terrorists). The process extracts high quality of information that can be used to draw conclusions based on relationships or pattern within the data.

In data mining Clustering learning is a popular and well researched way of discovering interesting results among huge database. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data pre-processing and model parameters until the result achieves the desired properties. There are many types of clustering techniques.

Typical cluster models include:

1. Connectivity models
2. Centroid models
3. Distribution models

4. Density models
5. Subspace models
6. Group models
7. Graph-based models

Among all the popular ways well verified and mostly used ways is K-means algorithm. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done. At this point we need to re-calculate k new centroids as center of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.
Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (1)$$

Here $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster center c_j, is an indicator of the distance of the *n* data points from their respective cluster centers.

1.1 Knowledge discovery in databases

Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery.

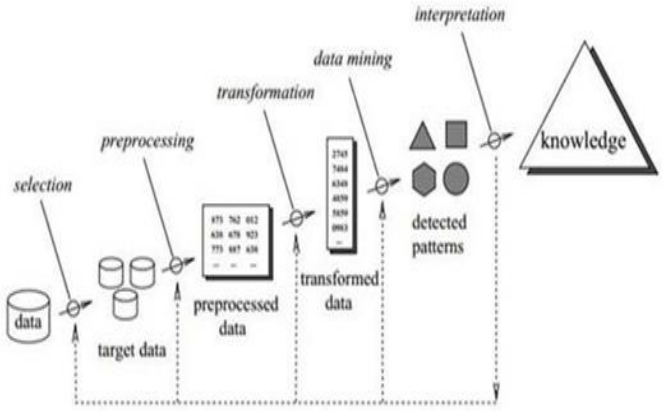

Fig 1.1: Knowledge Discovery in Databases

- **Data cleaning** - to remove noise and inconsistent data
- **Data integration** - where multiple data sources may be combined.
  A popular trend in the information industry is to perform data cleaning and data integration as a preprocessing step, where the resulting data are stored in a data warehouse.
- **Data selection** - where data relevant to the analysis task are retrieved from the database.
- **Data transformation** - where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
  Sometimes data transformation and consolidation are performed before the data selection process, particularly in the case of data warehousing. Data reduction may also be performed to obtain a smaller representation of the original data without sacrificing its integrity.
- **Data mining** - an essential process where intelligent methods are applied to extract data patterns.
- **Pattern evaluation** - to identify the truly interesting patterns representing knowledge based on interestingness measures.
- **Knowledge presentation** - where visualization and knowledge representation techniques are used to present mined knowledge to users.

1.2 MOODLE

A learning management system (LMS) is a software application for the administration, documentation, tracking, and reporting of training programs, classroom and online events, e-learning programs, and training content.
Moodle (Modular Object-Oriented Dynamic Learning Environment) is a course management system (CMS)

Roll no	Name	Q1/2.08	Q15/2.08	Q21/2.08	Q35/2.08	Q48/2.08	Grade/100
1	John	2.08	0	2.08	2.08	0	50.06
2	Mac	2.08	2.08	2.08	0	0	56.08
3	Steve	0	0	0	2.08	2.08	18.72
Overall avg.		18.72	0	31.2	31.2	2.08	

Table 2.1 Data sheet Overview

- a software package designed to help educators create quality online courses and manage learner outcomes. Such e-learning systems are sometimes also called Learning Management Systems (LMS), Virtual Learning Environments (VLE) and Learning Content Management Systems (LCMS). It is one of the most user-friendly and flexible open source courseware products available.

1.3 RAPID MINER

Rapid Miner, formerly known as YALE (Yet Another Learning Environment), is a software platform that provides data mining and machine learning procedures including: data loading and transformation (Extract, transform, load (ETL)), data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment Rapid Miner is written in the Java programming language. The internal XML representation ensures standardized interchange format of data mining experiments and the scripting language allows for automating large-scale experiments. Also it uses the multi-layered data view concept which ensures efficient and transparent data handling.

1.4 Problem definition

Study and analysis of K-means clustering algorithm using Rapid Miner for analyzing students' performance in examination and generating a way using Clustering Data mining technique to enhance students' performance. Purpose of clustering is to judge questions and on basis of that helping student to enhance their results.

II. Implementation

The aim of project was to find out a way to enhance student's performance for different category of students by using K-means clustering algorithm. To collect student exam data a Moodle test was conducted containing 48 questions. Each question was from different topic of major artificial intelligence. The timing given to exam was 50 minutes .This exam was attended by 95 students of same class. Each question was provided with 2.08 marks and grade was calculated on basis of that. Overall average was calculated by the formula:

$$\text{Overall avg.} = \frac{\text{Total correct attempts for a particular qn.} * 2.08}{\text{Total strength of class}} \quad (2)$$

The data sheet generated was alike as follows:

Later to divide questions K-means clustering was performed on Overall average (weight).Taking K=2 and K=3 and the results found are as follows:

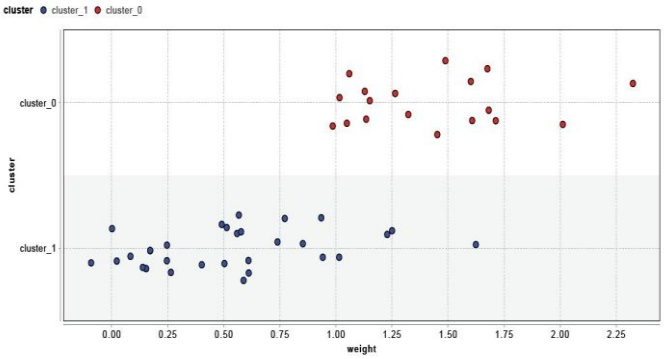

Fig 2.1: Clusters formed by K-means in Rapid Miner (K=2)

The questions divided by K-means are as follows:

Cluster 0 (30)		Cluster 1 (18)	
Question number	Weight	Question number	Weight
1	.42	3	0.96
2	.37	12	1.4
4	.59	13	1.38
5	.53	14	1.03
…	…	…	…
…	…	…	…
18	.53	48	1.70
Centroid	0.550	Centroid	1.329

Table 2.2: Cluster wise question division when K=2

And by analyzing the questions and their centroids clusters were named as:
Easy (Cluster 1)
Tough (Cluster 0)
Since this analysis is being performed on student's performance so it can be said that the ease and toughness of

Question is purely dependent on class performance. Or here definition of toughness is derived from class performance.

Later according to grades (total marks) students were divided into four categories and for a particular category average performance in cluster was calculated by using formula:

$$\text{Avg. performance of cluster} = \frac{\text{Total number of correct answers of cluster} * 2.08}{\text{Total number of students}} \quad (3)$$

Similarly for K=3 operations performed are as:

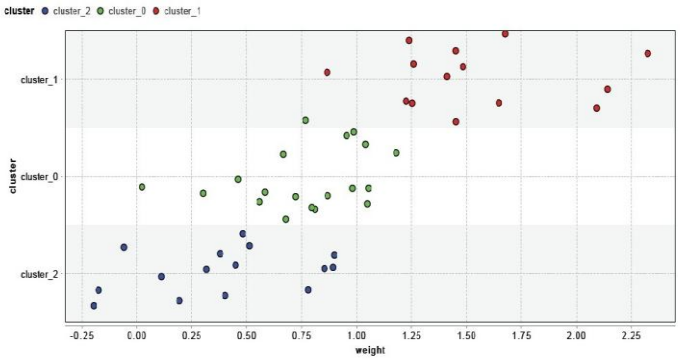

Fig 2.2: Clusters formed by K-means in Rapid Miner K=3

Here by analyzing the questions and their centroids clusters were named as:

Moderate (Cluster 0)
Tough (Cluster 1)
Easy (Cluster 2)

The questions divided by K-means are as follows:

Cluster 0 (20)		Cluster 1 (14)		Cluster 2 (14)	
Question	Weight	Question	Weight	Question	Weight
3	.96	1	.42	12	.14
4	.59	2	.37	13	1.38
7	.64	5	.53	16	1.32
10	.77	6	.35	19	1.12
11	.68	8	.42	22	1.14
…	…	…	…	…	…
…	…	…	…	…	…
46	.57	18	.46	48	1.71
Centroid	.7699	Centroid	.369	Centroid	1.418

Table 2.3: Cluster wise question division when K=3

And on this similar operations were performed like K=2

III. RESULTS AND DISCUSSIONS

Here an efficient and easy way is derived for enhancing students' performance. And by applying above implementation logics the results derived are the students having percentage above 75 are in A+ category having overall average between 45-60 total numbers of students are 28 in this category and their performance is 27.48 in cluster 0 and 21.76 in cluster 1 .Similarly in category A there are students having percentage between 60-75 there marks are between 36-44 total 45 students are there having average performance 23.85 in cluster 0, 15.53 in cluster 1.Whereas category B has students having percentage between 40-59 and marks between 24-35 there strength is 21 and cluster wise performance is 19.80 and 11.79 in cluster 0 and cluster 1 respectively. C marks less than 24 are in category C has students having percentage less than 40 there is only one student in this category having cluster performance as 10.4 and 8.32.

Here it is visible that students of whole class are weak in tough cluster so by applying clustering on questions of cluster1 recursively it can be found on which topics it is needed to work upon so as to enhance students' performance the validity of this process can be well seen by applying k=3. Here category A+ has performance greater than 75 marks 45-60 total students 28 having cluster performance as 20.28, 6.46, 22.50 in cluster 0, cluster 1 cluster 2 respectively ,whereas A has performance between 60-75 marks between 36-44 total students 45 having cluster performance as 14.83,4.89,19.64 in cluster 0, cluster 1 cluster 2 respectively, B has performance between 40-59 marks between 24-35 total students 45 having cluster performance as 1.49,3.96,17.13 in cluster 0, cluster 1 cluster 2 respectively. And C has performance less than 40 marks less than 24 total student 1 having cluster performance as 4.16, 6.24, 8.32 in cluster 0, cluster 1 cluster 2 respectively

Here data analysis comes out with more acceptable result as it is clearly visible that Students of category B are good in Moderate section that is questions of cluster0 as compared to Tough section that is Cluster1.And students of category C are good in Tough section as compared to their performance of Moderate section. So by working on students of category C with topics covered by Moderate section we can improve overall performance of students of category C.

Performance Index	0-60	No. of student	Performance Cluster0 Easy	Performance Cluster1 Tough
A+ >75	45-60	28	27.48	21.76
A 60-75	36-44	45	23.85	15.53
B 40-59	24-35	21	19.80	11.79
C <40	<24	01	10.4	8.32

Table 3.1 Student performance when K=2

Performance Index	0-60	No. of students	Performance Cluster0 Moderate	Performance Cluster1 Tough	Performance Cluster2 Easy
A+ >75	45-60	28	20.28	6.46	22.50
A 60-75	36-44	45	14.83	4.89	19.64
B 40-59	24-35	21	10.49	3.96	17.13
C <40	<24	01	4.16	6.24	8.32

Table 3.2 Student performance when K=3

IV. CONCLUSION AND FUTURE SCOPE

The result of this analysis gives clear idea about validity of this approach, this analysis was performed on a small dataset consisting of students and exam marks. It is well analyzed that this approach comes out with acceptable results. Hence this approach can be applied to other datasets having more complex features.

With developing data base and data analysis approaches data mining is becoming very common criteria hence in coming future it is going to be pioneer approach towards data analysis as in the short-term, the results of data mining will be profitable, in mundane, business related areas. Advertising will target potential customers with new precision. In the medium term, data mining may be as common and easy to use as e-mail. This tool can be used to find the best airfare, root out a phone number of a long-lost classmate, or find the best prices on lawn mowers. The long-term prospects are truly exciting. Imagine intelligent agents turned loose on medical research data or on sub-atomic particle data. Computers may reveal new treatments for diseases or new insights into the nature of the universe.

Big data is a term for a collection of data sets so large and complex that it becomes difficult to process using on-hand DBMS tools or traditional data processing applications. The challenges include capture, storage, search, sharing, transfer, analysis and visualization.

With developing IT and its uses data sets being developed are so large that they are so complex to be handled in coming future the analysis and Big Data is giving rise to an interesting collaboration among diverse disciplines of computer science, communication networks and devices, and behavioral science. The advent of data science as a mainstream subject is the outcome of these cross-domain efforts. Applying Big Data solutions, enterprises can now translate mountains of digital data into effective business insights in real time. They can avoid risks, cut costs and analyze patterns to follow trends and customers' preferences and suggest better choices for the customers and increase revenue.

REFERENCES

[1] A1-Sultan K. S.," A tabular search approach to the clustering problem, Pattern Recognition",pp-28:1443-1451,1995.

[2] Al-Sultan K.S. Khan M. M., "Computational experience on four algorithms for the hard clustering problem",pp-295-308,1996.

[3] Banfield J. D. and Raftery A.E.,"Model-based Gaussian and non-Gaussian clustering",pp-803-821, 1993.

[4] Bentley J. L. and Friedman J. H.," Fast algorithms for constructing minimal spanning trees in coordinate spaces.", IEEE Transactions on Computers, C-
pp-297-105, February 1978. 275

[5] Aggarwal, C.C., Hinneburg, A.2000,"On the distance metrics in high dimensional space", IBM Research report, RC 21739. Aggarwal, C.C., procopiuc, C., Wolf, J.L., YU, P.S., J.S. 1999a.,"Fast algorithms for projected clustering",In Proceedings of the ACM SIGMOD Conference, 61-72, Philadelphia, PA.

[6] Liu, X. and Croft, W. B. (2004),"Cluster-based retrieval using language models,"pp-234-256.

[7] Wenceslas Fernandez de la Vega, Marek Karpinski,Claire Kenyon, and Yuval Rabani. Approximation schemes for clustering problems. In Proceedings of the 35th Annual ACM Symposium on Theory of Computing (SIGIR '03), pages 50–58, 2003., New York, NY, USA. ACM.