# Patterns in Teaching Staff Recruitment in India (2015)

Piyush Kalantri Inala Vivek Vamsi Karri Srikanth Saksham Gupta

2015B4A70529H 2016AAPS0230H 2016AAPS0234H 2018A7PS0281H

*Abstract*—Data Mining is a process to extract information for developing significant relationship with variables stored in large data warehouses. Education is an essential element for the progress of country. Educational data mining is a discipline processes to developing methods for exploring the unique type of data from educational settings. Those methods are used to better understand students and the settings which is defined by the educational data mining. Mining in educational environment can processing data is called educational data mining. Developing new methods to knowledge discover process from educational databases systems. This educational data mining provides a set of techniques to help educational system. The objective of this research is to introduce educational data mining in describe step by- step process using technique of K-means Clustering methods. In this evaluation factors students are like to write mid-term exams and final exams assignments. This study will use to help the teacher to reduced drop-out ratio to a significant level and improving the performance of students.

Keywords—Data Mining, Educational Data Mining, K-Means, Clustering, Higher Education System.

I. Introduction

Data mining is a process to discover knowledge from data which the stored in data base and data warehouses responsibility. Educational system face several issues may be including students marks like pass, fail and percentage to identify their data. Educational data mining (EDM) system used to improve the students marks in the several techniques like clustering. The mining process is use to educate or identify a student in the high, medium and low. This k-means clustering method uses process to develop the student mark and percentage in centred values. Data mining in higher educational system denote the mining process of data in a data warehouses like hierarchies, partition categories, sampling large data bases. Grade system requires the high, good, poor and low, medium. Education system uses a issues to identify students need, performance and training of quality interactions. EDM is a process used to increase the student percentage and profits. Also improve the learning process in mining data. Large data analysis is a manual process to automatically provide the student performance in a data to provide a profit values. User to allow the data from different dimensions relationship during mining processes

1. Data Mining Higher Education System

Education is an essential process for the better and progress of a country. It enables the people of a country civilized and well manners. Mining in educational environment is also known as Educational Data mining, with developing new methods to discovering knowledge from educational databases in order to analyze student’s trends and behavior in mining towards education. Lacks of data deep and enough knowledge in higher educational system may prevents system management to achieve quality objectives, data mining methodology can help bridges this knowledge gap in higher education system.

***B) Data***

Data is based on the survey conducted by the MHRD.The main objectives of the survey was to (1) - identify & capture all the institutions of higher learning in the country, (2) - Collect data from all the higher education institutions on various aspects of higher education. This Catalogue contains the data regarding Basic Information, Hostel Information, Departments, Non-Teaching Staff Details, Teaching Staff Details, Teaching Staff Summary, Student Enrolment for Regular Courses, Examination Results, Infrastructure, Fellowships, Loans, Accreditation, Scholarships, Other Minority Colleges, Other Minority Standalone Institutes and Other Minority Universities etc. for various Colleges, Standalone Institutes and Universities***.***

The Teaching Staff dataset contains information of technical staff recruits in the year 2015 across different universities and colleges in India. The dataset includes the college name, college id, department, designation of the recruits and the categories they belong to (namely General, OBC, ST, SC) along with their gender. There is a one-one mapping between college names and id, hence college names are redundant and is therefore ignored. College Id, department, designation and category are the four(4) attributes which we will consider for performing data mining tasks as mentioned in the Problem definition below.

C) Preprocessing

Data pre-processing is crucial in any data mining process as they directly impact success rate of the project. This reduces complexity of the data under analysis as data in real world is unclean. Data is said to be unclean if it is missing attribute, attribute values, contain noise or outliers and duplicate or wrong data. Presence of any of these will degrade quality of the results.

Here are few important data pre-processing techniques that can be performed are handling missing data, dimensionality reduction, feature selection, feature creation, binarization.

***D) Visualisation***

Without the concept of visualization, mining and analysis doesn’t play any role of importance as data mining is the idea of finding inferences by analysing the data through patterns and those patterns can only be represented by different visualization techniques.

## **Uses of data visualization**

* Powerful way to explore our data with presentable results.
* Primary use is the pre-processing portion of the data mining process.
* Supports in data cleaning process by finding incorrect and missing values.
* For variable derivation and selection means to determine which variable to include and discarded in the analysis.
* Also play role in combining categories as part of the data reduction process.

## Techniques:

* Box plots
* Histograms
* Heat maps
* Charts

## **II.Data Preprocessing**

The preprocessing techniques used here are :

1. ***Data Cleaning***

The **missing values** in each numerical attributes is replaced by **mean** of those columns. The categorial attribute is replaced by **mode**. Since the number of missing values in each column is significant , replacing with average value is more appropriate based on the assumption that the values can be closer to mean. This helped to increase data points so that it can be helpful to make better predictions.

1. ***Data Reduction***

The attributes which convey same meaning will not help for any data mining techniques so removing such features help in **dimensionality reduction** thus contributing towards feasible processing of data for pattern recognition. Attributes which do not contribute any significant information to our data were removed (including attributes names: survey\_year, id, grade\_pay, faculty\_name).

1. ***Data Transformation***

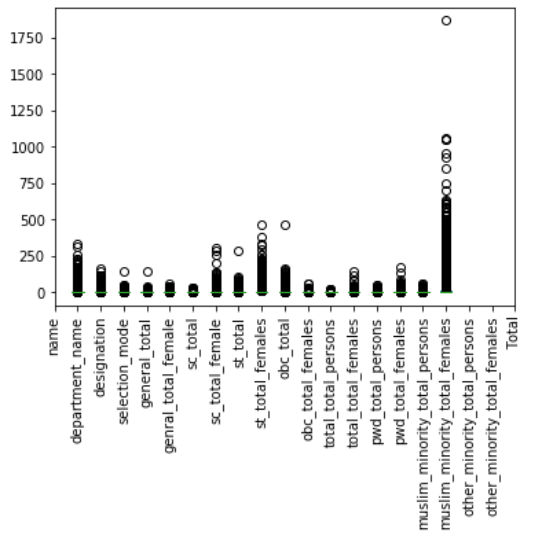
A new attribute was introduced **(feature creation)** which consists of the aggregate of multiple columns thus giving us an overview of data in each row. **Feature selection** was done thus redundant features were removed. **Binarization** technique was used on features having binary attributes.

## **III.Data Visualisation**

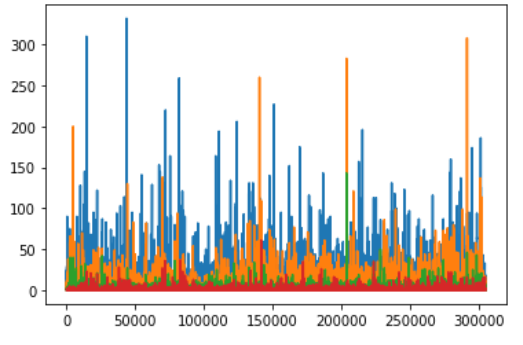
## 

Data was visualised using matplotlib, seaborn, apache superset, where pie charts, bar graphs, heat maps, plot box, pair plot, scatter matrix and word cloud were plotted between various attributes. Snippets of some of those graphs are shown below:

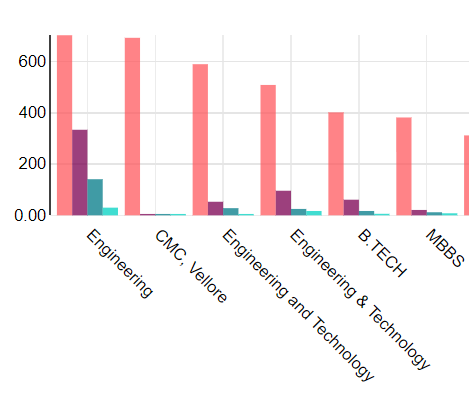
* **Box Plot showing outliers**



* **Category wise distribution of members**



#### **Bar Graph- Department wise category**



***IV. Techniques for obtaining results***

There are very good number of techniques in data mining to obtain the required results depending on the nature and characteristics of data. The techniques used here are Apriori and K means clustering. The Apriori is one of the association mining rule and K means is clustering

**A. *Associate Rule Mining***

Association rule mining finds interesting associations

and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. The Association rule is very useful in analyzing datasets. The data is collected using bar-code scanners in supermarkets. Such databases consists of a large number of transaction records which list all items bought by a customer on a single purchase. So the manager could know if certain groups of items are consistently purchased together and use this data for adjusting store layouts, cross-selling, promotions based on statistics.

The data here contains very of spectrum of colleges , professions, roles, categories. Its quite interesting to draw patterns here. The association rule mining is based on pruning the itemsets which are not frequent. This would make the patterns obtain easily and faster. The method starts from picking the 1-itemset and finding the frequent 1-itemsets and then finding the 2-itemsets from these frequent itemset and so. At every level we draw the frequent item sets. The rules can be generated from these frequent itemsets.

The association rule mining requires scanning the data everytime to find the support of each itemset. The data can be easily and read faster if the data is converted as market basket .

The support and confidence we set for out are 0.5,0.5 respectively

The frequent itemsets found from our data are :

({'st'}), ({'obc'}), ({'sc'}), ({'muslim\_minority'}), ({'general'})

Temporary Teacher'}), ({'Lecturer (Senior Scale)'}), ({'Part-Time Teacher'}), ({'Principal'}), ({'Ad hoc Teacher'}), ({'Demonstrator'}), ({'Lecturer (Selection Grade)'}), ({'Contract Teacher'}), ({'Tutor'}), ({'Reader'}), ({'Visiting Teacher'}), ({'Professor & Equivalent'}), ({'Lecturer'})

The patterns are:

({'Tutor'}) -> ({'General}) conf: 0.5317691029900333

({Muslim category}) -> ({'Assistant Professor'}) conf: 0.5448504983388704

({'Reader'}) -> ({st}) conf: 0.583351636283769

({'Associate Professor'}) -> ({'sc}) conf: 0.6563829310885237

({'Associate Professor'}) -> ({'st}) conf: 0.6563829310885237

({'Professor & Equivalent'}) -> ({obc'}) conf: 0.7546968403074297

({'Professor & Equivalent'}) -> (`st'}) conf: 0.596968403074295

({'Professor}) -> (‘obc’) conf: 0.5401793339026474

({‘obc'}) -> ({'Assistant Professor'}) conf: 0.6563829310885237

({'Professor & Equivalent'}) -> ({‘general'}) conf: 0.7546968403074297

({‘reader'}) -> ({‘muslim category'}) conf: 0.5969684030742954

These are some of the patterns derived from the data with their respective confidence values.

## **B.K-means Clustering:**

**Clustering** is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

Clustering analysis can be done on the basis of features where we try to find subgroups of samples based on features or on the basis of samples where we try to find subgroups of features based on samples. We’ll cover here clustering based on features. Clustering is used in market segmentation; where we try to fined customers that are similar to each other whether in terms of behaviors or attributes, image segmentation/compression; where we try to group similar regions together, document clustering based on topics, etc.

Unlike supervised learning, clustering is considered an unsupervised learning method since we don’t have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance. We only want to try to investigate the structure of the data by grouping the data points into distinct subgroups.

Unlike supervised learning, clustering is considered an unsupervised learning method since we don’t have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance. We only want to try to investigate the structure of the data by grouping the data points into distinct subgroups.

**Kmeans** algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

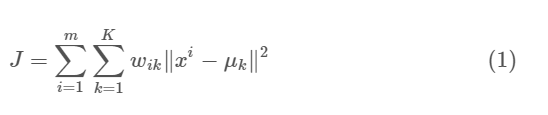
The way kmeans algorithm works is as follows:

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K*data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.

* Compute the sum of the squared distance between data points and all centroids.
* Assign each data point to the closest cluster (centroid).
* Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

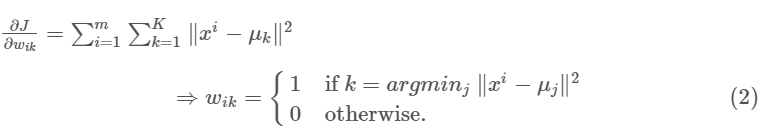
The approach kmeans follows to solve the problem is called **Expectation-Maximization**. The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster. Below is a break down of how we can solve it mathematically.

The objective function is:

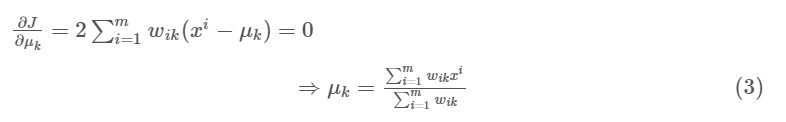


where wik=1 for data point xi if it belongs to cluster *k*; otherwise, wik=0. Also, μk is the centroid of xi’s cluster.

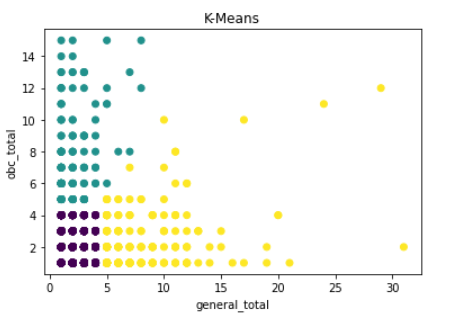
It’s a minimization problem of two parts. We first minimize J w.r.t. wik and treat μk fixed. Then we minimize J w.r.t. μk and treat wik fixed. Technically speaking, we differentiate J w.r.t. wik first and update cluster assignments (*E-step*). Then we differentiate J w.r.t. μk and recompute the centroids after the cluster assignments from previous step (*M-step*). Therefore, E-step is:



In other words, assign the data point xi to the closest cluster judged by its sum of squared distance from cluster’s centroid.And M-step is:



The plot of K means clustering of our data is found and is below:

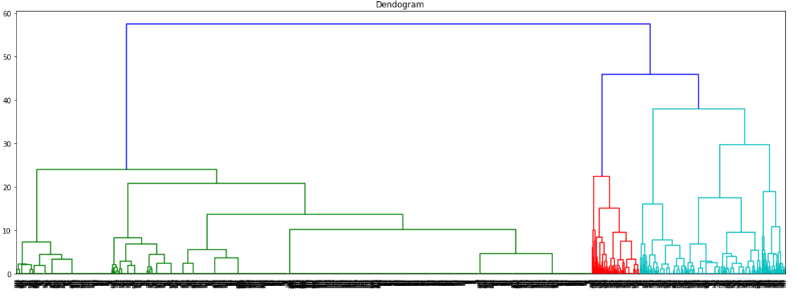


## Dendograms

A dendrogram is a diagram that shows the hierarchical

relationship between objects. It is most commonly created as an output from [*hierarchical clustering*](https://www.displayr.com/what-is-hierarchical-clustering/). The main use of a dendrogram is to work out the best way to allocate objects to clusters

The dendograms for our data is given below



***IV. Conclusion***

The preprocessing of the data is done based on various techniques to make it more meaningful and useful. The techniques helped the data to become clean and more expressive.

It also helps in making better models and data mining techniques to give meaningful. Visualization is very useful , it helps in giving a vague idea of whole data. The various visualization tools helps in presenting data more pleaseful.

***V. Acknowledgements***

We would to like to express my appreciation and gratitude for the lectures taught by Prof. Manik Gupta. The lessons were quite informative and helped in completing this assignment. Also we would like to thank all the Teaching Assistants for helping all the time throughout the assignment.

***V.References***

[1] The data is taken from [https://data.gov.in](https://data.gov.in/)

[2] Techniques taught in class.