Data Mining on University Accreditation

Using Python

Aman Syed
Birla Institute of Technology and Science Pilani, Hyderabad Campus
f20160046@hyderabad.bits-pilani.ac.in

Archishman Gupta
Birla Institute of Technology and Science Pilani, Hyderabad Campus
f20160051@hyderabad.bits-pilani.ac.in

Gaurav Hada
Birla Institute of Technology and Science Pilani, Hyderabad Campus
f20160582@hyderabad.bits-pilani.ac.in

Harish Goud Ediga
Birla Institute of Technology and Science Pilani, Hyderabad Campus
f20160110@hyderabad.bits-pilani.ac.in

***Abstract*—The system proposed in the paper is an advanced solution for extracting insights form the given raw data from a government survey. The data mining technique used in the project not only refines the data but also helps in extracting useful information from the data. The project deals with data selection from the pool of datasets by understanding the data in detail which involves identifying the key questions or insights for the chosen data. Post finalizing the dataset, performing exploratory data analysis along with suitable visualizations using Power BI and employing different prepossessing techniques using python which are suitable for the dataset. After completion of data prepossessing techniques, based on the insights needed from the data, identifying the key Data Mining tasks like association analysis, clustering, classification or outlier analysis on the data with python algorithms which will help in achieving the results.**

I. Problem Motivation

The available data on Indian Universities provide a lot of opportunities for various sorts of deductions. But the data is present in a raw format which contains a lot of meaningless values and noise. In order to deduct something out of it, the data needs to scrub or cleansed and pre-processed. Only then, is the data fit for further processing and deduction. Also, there is an ardent need to classify or sort the data into meaningful cluster, to aid in providing further insights into it. These deductions and analysis might be of assistance to the Indian Government in future.

II. Background

The datasets for our use were picked from https://data.government.in. This data is an assimilation from the surveys conducted in 2015 on different Indian universities from various backgrounds.

III. Objectives

The Government of India initiated a survey program in order to identify the accredited universities and to maintain a record of their respective scores. In order to get useful insights from the data, the team initiated this project employing data mining on the selected datasets in order to achieve the following objectives:

- To perform data cleansing task without the manipulation of important information.
- To divide the data into clusters to provide meaningful data insights.
- To find out the number of universities which are rated above average in India.
- To find out the distribution of universities with respect to the highest achievable score.
- To find out the number of universities which have been accredited by various bodies.

IV. Introduction

When raw data is obtained for processing, it is usually in an impure form, i.e., it may contain empty or missing values, out-of-bounds values (for ex: In a column which is fit for positive integers, the value '-1' may be present) or even values from a different data type (for ex: In a column fit for integer type data, the data 'six' may be present). This data is yet not ready for processing. However, applying cleansing techniques remove these anomalies and render the data fit for further processing. The objective of data cleansing is not just to clean up the data in a dataset, but also to bring about consistency to the different sets of data that have been merged from separate datasets. Furthermore, we have applied K-means clustering to group the data elements into their related groups. The objects are grouped based on the information found in their description, or on a particular relationship. The goal is that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The following report is an attempt to briefly discuss the various steps involved in this particular data analysis.

V. Data Aggregation

Data aggregation is any process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis. A common aggregation purpose is to get more information about groups based on specific variables such as age, profession, or income.

VI. Discretization

Discretization refers to the process of converting or partitioning continuous attributes, features or variables to discretized or nominal attributes/features/variables/intervals. For example, attribute values can be discretized by applying equal-width or equal-frequency binning, and then replacing each bin value by the bin mean or median, as in smoothing by bin means or smoothing by bin medians, respectively. The continuous values can be converted to a nominal or discretized value which is same as the value of their corresponding bin.

VII. Normalization

In simple words, when the data consists of multiple attributes but these attributes have values on different scales, this may lead to poor data models while performing data mining operations. So they are normalized in order to bring back to the same scale. Normalization simply means transforming the data, converting it into another format to minimize the range of data, thus aiding in efficient data processing. It is a data pre-processing step. The following are the different methods of data normalization.

A. Decimal Scaling Normalization

It normalizes by moving the decimal point of values of the data. To normalize the data by this technique, we divide each value of the data by the maximum absolute value of data. The data value v, is normalized to v', by using the formula below

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $\max \mid v' \mid < 1$

B. Min-Max Normalization

In this technique of data normalization, linear transformation is performed on the original data. Minimum and maximum value from data is fetched and each value is replaced. Let (X_1, X_2) be a min and max boundary of an attribute and (Y_1, Y_2) be the new scale at which we are normalizing, then for v value of the attribute, the normalized value v' is given a

$$v' = \frac{v - X_1}{X_2 - X_1}(Y_2 - Y_1) + Y_1$$

C. z-Score Normalization

In this technique, values are normalized based on mean and standard deviation of the data A. The formula used is

$$v' = \frac{v - \mu}{\sigma}$$

where μ is the mean value of the feature and is the standard deviation of the feature

VIII. Clustering

The first step in data pre-processing is that of clustering. A cluster is a collection of data points, and clustering is the process of dividing a dataset into finite clusters, such that data points in the same cluster are similar (or related) to each other and are different from (or unrelated to) those in other clusters. Clustering is very important as it is an unsupervised learning method i.e. it determines intrinsic grouping among unlabelled data. There are various types of clustering methods, listed as follows.

A. Density-Based Methods

These methods consider the clusters as the dense region having some similarity and different from the lower dense region of the space. These methods have good accuracy and ability to merge two clusters.
Examples: DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure)

B. Hierarchical Methods

The clusters formed in this method forms a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two categories

- Agglomerative (bottom-up approach)
- Divisive (top-down approach)

Examples: CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies)

C. Partitioning Methods

These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter.
Examples: K-means, CLARANS (Clustering Large Applications based upon Randomized Search)

D. Grid-Based Methods

In this method the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operation done on these grids are fast and independent of the number of data objects.
Examples: STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest)

In this case, we have used K-means clustering, to sort our dataset into 3 distinct clusters.

IX. K-Means Clustering

K-Means 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. The basic algorithm is as shown below

1) Select K points as initial centroids
2) **repeat**
    a) Form K clusters by assigning all points to the closest centroid
    b) Recompute the centroid of each cluster
3) **until** The centroids do not change

The approach K-Means 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

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

where SSE is the sum of sqaured errors, x is a data point in cluster C_i and m_i is the representative point for C_i

We have set K as 3 for our dataset.

X. Methodology

While working on the University Accreditation data, we observed that the crude .csv files contain missing values, redundant data and inconsistent values (change in measure scales like A,B,C and 1,2,3). So, we have identified the most important attributes and removed the rows with null values. Then, we replaced the missing values in other columns with mean/probable value depending on other columns. After that, we proceeded to perform normalization so as to make the data consistent, followed by Attribute subset selection in which we removed the columns which are consistent and have the same value all over the data, for ex. survey year. Following that, we removed the columns which are redundant like college ID and college name, which are unique. We then proceeded to Discretization by replacing the unique attributes in the column with a corresponding numerical value by using encoding from scikit-lean library so as to effectively perform exploratory analysis. After that we proceeded to Aggregation as we could see few columns interpreting a similar trend. Then we summed up all these values and made a new column for two such similarities. After following all these procedures, we ended up with data which is filled and has attributes which has co-relations that can be found in later part of the project.

From the cleaned data, various columns as parameters were considered, and were converted into respective lists. Now, the plot was made out of the pairing of two relevant distinct lists. According to the plot, the number of clusters was identified and using the K-Means clustering algorithm, the value of k was substituted by the number of clusters. In this process, the random centroids of the k clusters were calculated and represented on the plot by means of an '*' (asterisk) symbol. Then, the plot points were sorted in accordance with the closest cluster. To identify the points of the K clusters, they were colored distinctly which provides better visualization.

Finally, for visualization, we plotted the data using pyplot from matplotlib and visualized it in the form of scatter plot and bar graph.

XI. Results

From the given data we were given initially, we managed to cleanse the data and reduce any unnecessary attributes previously present. After applying normalization and other such techniques, we then moved on to clustering the data into meaningful groups so as to gain insights into the data. We arrived at a few insightful plots from our analysis which we have added into the Plots folder as we were unable to attach it over here on LaTeX.

XII. References

- https://github.com/aman3599/CSF415-Team-26.git