EasyChair Preprint

№ 6821

Extracting Insights by Clustering Structured Data

Amir Hossein Rouhi

October 9, 2021

Extracting Insights by Clustering Structured Data

Amir H. Rouhi

Data and Analytics, Finance and Governance, RMIT University, Melbourne (amir.rouhi@rmit.edu.au)

ABSTRACT

As part of the higher education ecosystem, Institutional Research (IR) is an integral part. Institutional data is one of the building blocks that makes IR vital in decision making and shaping policy and strategy. All the institutional entities —. students and courses consisting of different attributes, such as program code and name, course code and names, credit points, etcetera — are stored in defined structures named tables. These tables are conventionally stored in the form of structured data elements (fields or columns) and tuples (records or rows) in Relational Database Management Systems (RDBMSs). Breaking down the concept of entities and their attributes and storing them into tables is called normalization. This process is for reducing the data redundancies which is the main concern in large RDBMSs. Hence, given the fact that the entities and their attributes are the concepts already categorized and stored in the database tables, to what extent can this cliché structure negatively impact on researchers by limiting their views to the institutional data?

The objective of this research presentation is to introduce a new lens by which to analyze structured data with the aid of Clustering algorithms. To achieve this objective, the attributes of different entities can be merged using classical database views. Before we embark on the conventional analysis of the extracted data, we can apply an unsupervised Machine Learning algorithm (Clustering) to detect hidden correlations among the attributes and thereby re-group the datapoints into new clusters in order to start the analyzing process. This can assist institutional researchers to distill different perspectives of data and to extract invaluable insights based on the automatically detected clusters. The key factor in this approach is defining the appropriate number of clusters and, subsequently, the interpretation skills for the new clusters.

Keywords: Machine Learning, Clustering, Insight extraction, Structured data, RDBMS.

Introduction

The education sector, like any other organization, necessarily utilizes relational databases to store daily transactional data into pre-defined structures, known as tables. Data forms the building blocks of all computer-based systems. All these systems are the product of primary requirement analysis, such as structured-based (SSADM) or object-oriented-based (OOAD) software engineering processes. Regardless of the analyzing methods, the analyst focuses on the process or objects of the system and naturally categorizes conceptually correlated attributes together, which will ultimately form the final tables in the databases. This process will divide the concepts into entities and attributes and store them is different database tables. Tables are connected to each other based on their key attributes, so the analyst will be able to connect the tables to extract more complicated concepts in form of database views or even output reports.

The benefits of databases are obvious; almost no business can function without utilizing them nowadays. The amount of transactional data generated in each hour or day is beyond classical data storage capabilities. Moreover, to ensure adequate data storage capacity, the complicatedness of requests needed to run a successful business forces the utilization of databases by designing views and reports. Briefly, improving business management is the byproduct of computerized management systems and their databases.

As explained above, analyzing business, and dividing the business concepts into entities, and entities into correlated attributes, helps to conquer the difficulties and ambiguities in business management. However, neither the dividing processes nor the forging of entities' attributes into defined tables guarantees that all the correlations between attributes has been captured during the process of system analysis. There is always a possibility of the

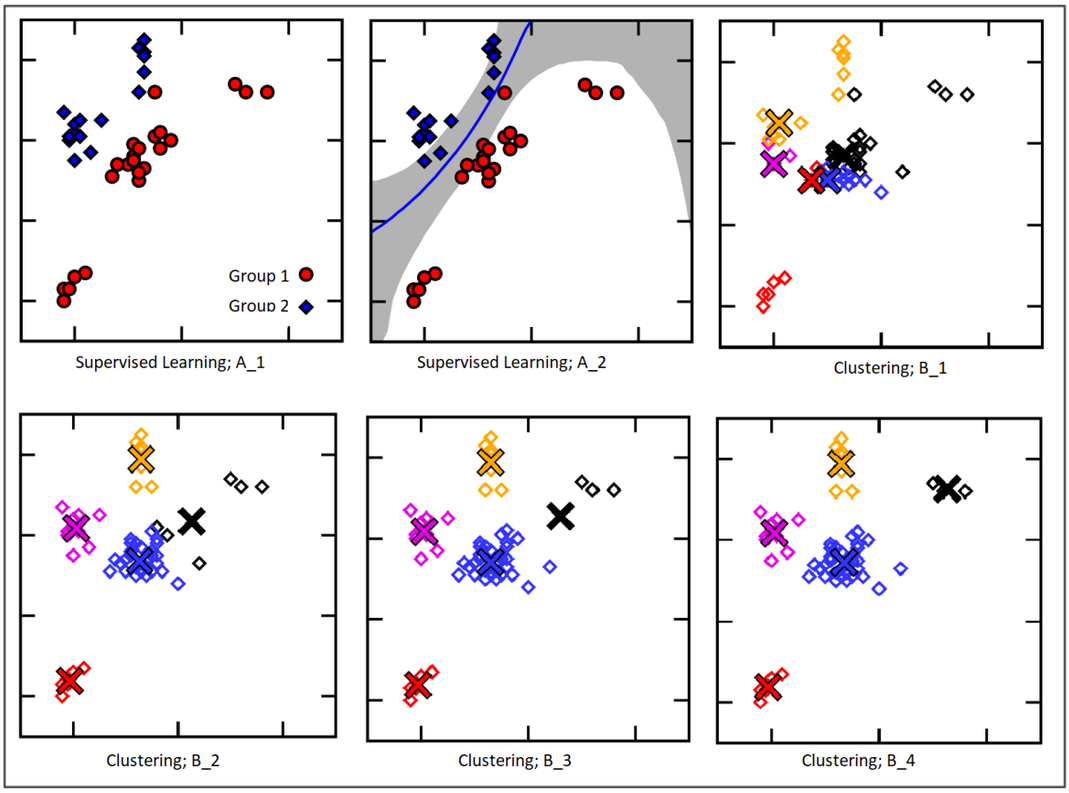

Figure 1: Supervised versus unsupervised learning. A_1 and A_2 represent supervised learning and groups of data are known. B_1 to B_4 illustrates the Clustering process as an unsupervised learning. As can be seen a randomly selected centroids and groups of data in B_1, finally ended with nicely clustered data in B_4.

existence of unknown correlations among attributes in the same table, or in different tables of a database, or even in different databases of a data warehouse.

Fortunately, there are some tools and techniques that allow the investigation and extraction of such hidden correlations or patterns. Unsupervised learning is one Machine Learning method that helps to categorize stored data beyond their technical database structures and systems. In this research, the way in which Clustering as an unsupervised learning tool helps to distill such patterns or categories, and enriches our knowledge of our business, is demonstrated.

After this short introduction, the following sections are provided in this research:

- Introduction to types of Machine Learning and Clustering
- How Clustering helps to extract insights?
- Applying Clustering on institutional structured data
- Conclusion

Introduction to Machine Learning and Clustering

In Artificial Intelligence (AI) and Machine Learning (ML), there are 3 main paradigms for the learning rule: Reinforcement (RL), Supervised (SL) and Unsupervised Learning (UL). The former two paradigms are core methods widely used in different applications (Ayodele, T.O., 2010). The main difference between the two is the utilization of labeled date in SL and unlabeled data in UL. The information in the training data for RL is intermediate between SL and UL (Jordan, M.I., 2015).

The algorithms of SL needs labeled data to map the input to the labeled output. The SL process adjusts the weight parameters of numerous functions in different layers (input, middle and output) in a way that map the input to the desired output (Jordan, M.I., 2015). This process happens in the learning phase and when the system is trained on all the labeled data, it is ready for the predicting phase to automatically map any unknown input to the output (Figure 1; A_1 and A_2). The more appropriately the data for the training phase is selected and labeled, the greater the accuracy of the system in prediction phase. There exist

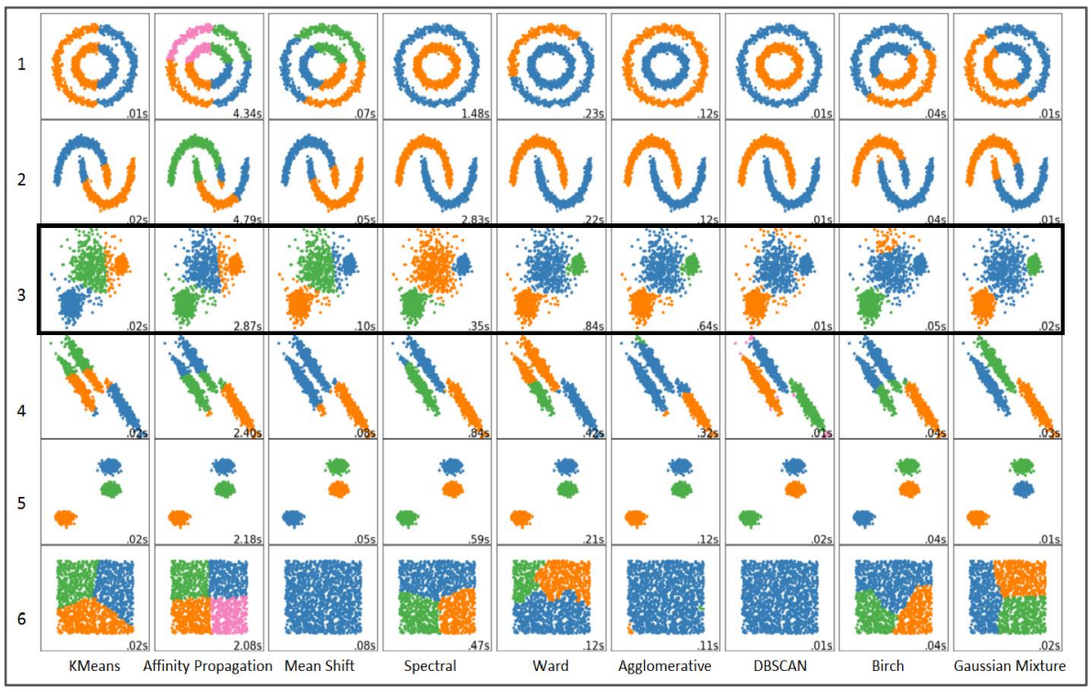

Figure 2: Clustering algorithms. As can be seen the formation of data distribution is important to select appropriate algorithm.

different types of mapping functions $f(x)$ in SL which generate an output y for input x. Some of the well k`12345-+
0-`1238n0

own functions are neural networks, decision trees, decision forests, logistic regression, support vector machines and Bayesian classifiers (Hastie, 2011). The SL models are widely used in classifications and regression problems.

Unlike SL, the algorithms of UL do not need labeled data to map input to output for the training phase. Their algorithms automatically investigate the data, based on assumptions of the structural properties of the data, to discover inherent patterns or structures (Jordan, M.I., 2015). However, they need some input parameter such as the number of clusters (k) in Clustering algorithms. They also need human interpretation to validate their outputs (De Lua, 2021). The three main tasks for ULs are Clustering (i.e. *k-means* data grouping), Association (i.e. market analysis), and Dimensionality Reduction (i.e. topic modeling).

In some applications, both SL and UL are employed together. When the datasets are huge and labeling data manually is almost impossible, Clustering and Dimension Reduction can be utilized for automatically labeling datapoints to make them available for SL.

Clustering can be known as the art of detecting implicit knowledge in the absence of explicit labels, which can support the grouping of datapoints into clusters. There exists a wide range of Clustering models, such as Centroid models (*K-means*), Connectivity models (*Hierarchical Clustering*), Density-based Clustering (*DBSCAN*) and *Affinity propagation*, which can be variously selected based on the nature of the "Cluster" in the application and datapoints.

Due to the pattern of data distribution in the current research, *K-means* has been employed as the selected Clustering algorithm. *K-means* is a model-based, centroid model Clustering algorithm and its properties makes it the most popular Clustering algorithm. Generally, it can be applied on a wide range of Clustering problems. Its algorithm represents each cluster by a single mean vector. In this algorithm, the number of clusters (classes, groups) needs to be selected and the algorithm initializes by assigning random center-points for each randomly selected group. Choosing the number of groups is experimental, and the selection is made heuristically or based on experience or on the application's constraints. Each datapoint is classified by its distance from the center point (centroid), which is calculated by a distance function i.e. Euclidean. Based on the mean distances of the datapoints from the random centers, the new centers will be re-computed and the process of calculating the mean distances from the new centers will be repeated. These steps will be repeated in several iterations until the mean distances from the group centers do not change

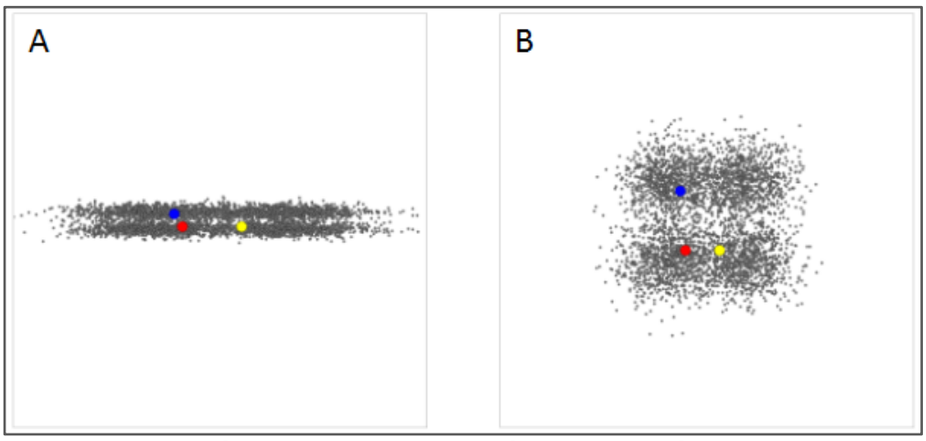

Figure 3: The impact of normalization; A- before and B- after normalization on the same dataset.

significantly (Figure 1; B_1 to B_4). The cluster labels on the datapoints in this status are interpreted as the most appropriate Clustering. *K-means* is a very efficient algorithm and selecting the number of groups is not always trivial, because the objective is to extract insight from the data. *K-Median* is another version of *K-means* which is less sensitive to outliers, but computationally more expensive. Figure 2 illustrates different Clustering algorithms (Scikit Org.). As can be seen, the distribution pattern of datapoints is the key factor in selecting the Clustering methods. The data distribution pattern used in this research is more like the form of distribution in the 3rd row. As can be seen, the results for *K-means* are exactly like the other two algorithms: *Affinity Propagation* and *Mean Shift*, and do not display a significant difference to those of the other algorithms.

If N represents total datapoints and X_n represents each of them and k represents the number of clusters and m_k represents the centroid of the cluster, the cost function for the *K-means* algorithm is as follows:

$$C = \sum_{n=1}^{N} \sum_{k=1}^{K} S_{nk} \, \|X_n - m_k\|^2$$

where $S_{nk} = 1$ if data point: n is assigned to the clusters: k and $S_{nk} = 0$ otherwise. It's important to know that $\sum_{k=1}^{K} S_{nk} = 1$, which means a datapoint can be assigned to one cluster only. The objective in the *K-means* algorithm is to minimize C.

The process of utilizing Clustering in extracting insight

In the previous section, the way in which the *K-means* algorithm can help to cluster datapoints into groups automatically is explained, based on minimizing the cost function. The objective of this section is to describe how this ability can be utilized on the structured data of institutional databases in order to find the hidden correlation among datapoints (attributes of entities) and to finally utilize it to distil new insights.

The first step is defining a problem. The objective in the problem statement should be realistic and in harmony with the maturity of data in our institutional databases. The way in which the problem is defined in undergraduate programs will be explained in the following sections.
The next step, a technical one, is related to extracting structured data from the databases. If all the attributes of the needed data are already recognized as related attributes of an entity, it is possible to extract the datapoint from a single table of one of the databases. However, in most of the problem statement, different aspects of entities need to be combined, before any Clustering phase, into one data extract. In such cases, a View to extract data from different tables in a database or other databases in the data warehouse needs to be designed. In either case, the output of this phase of data extraction from the structured data is a table or worksheet, in which it is expected there will be some pattern correlations among the datapoints; such correlations are the subject of interest.

Preprocessing the data before applying the Clustering algorithm is almost essential. The type of data, and the way in which they are stored in databases, is not necessarily appropriate for Clustering. The most common-preprocessing activity is the normalization of data. This process helps to segregate the clusters more clearly; otherwise the distances between datapoints are not following the same standard and cannot be compared to each other. The normalization formula is as follows:

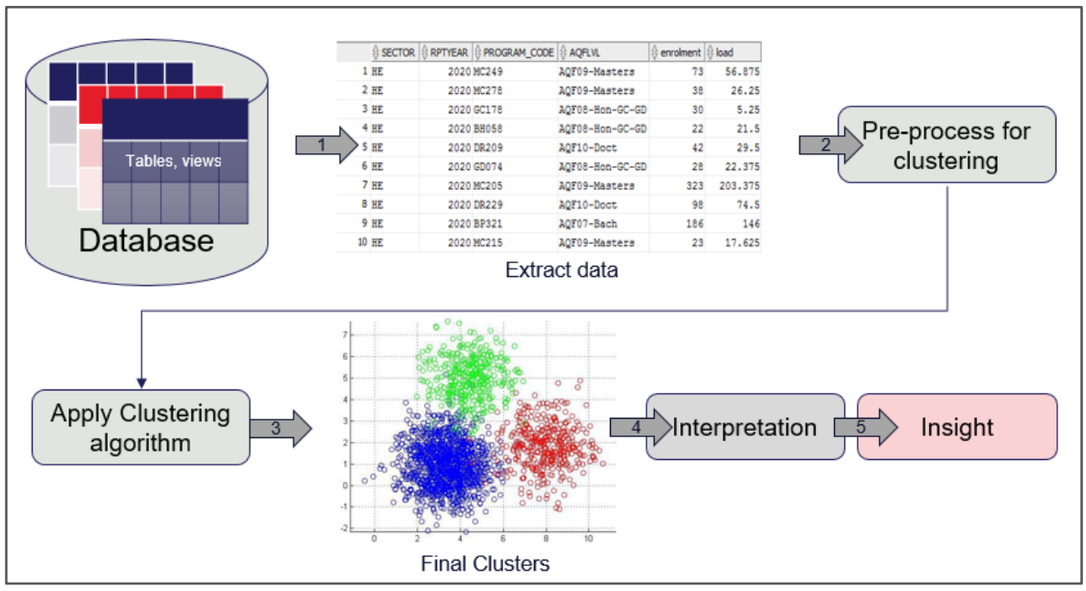

Figure 4: Process of Clustering structured data.

$$X` = (X - \mu)/\sigma$$

where $X`$ is the normalized X and $\mu = mean$ and $\sigma = standard\ deviation$. *Log Transforms* and *Quantiles* are also other techniques that can be used in data preparation; however, they are not utilized in this research. Figure 3 represents the Clustering and their centroids before and after normalization.

The second consideration in data preparation is related to the type of data. Clustering algorithms are designed for numerical data because it is needed to calculate the distance between the datapoints and the centroids. However, it is very common that we have some categorical (non-numeric) data among our datapoints, such as level of education (PGRD, UGRD) or results (Pass, Fail). There are some techniques to overcome this issue in Clustering. *K-modes* is among the first technique introduced by Huang which is based on dissimilarity measures to deal with categorical objects (Huang, Z., 1998). There exist other techniques, which are introduced in Potdar (2017); of these, the *Ordinal* and *One Hot* are easy to implement and are accurate encoding techniques. Both are utilized in this research.

After data preparation, the Clustering algorithm (*K-means*) can be applied on the data and the result will be ready for interpretation. It is possible to investigate and compare the results with a different number of clusters (K) to find the most meaningful number of clusters for the project. Moreover, to this heuristic approach, there are some techniques that are helpful in selecting the appropriate number of clusters. *Bayesian Information Criterion (BIC)* is a method that is often used in model-based Clustering; however, it can also be used in partitioning-based Clustering (Zhao, 2008). There is another method, known as *Kluster* procedure, which provides more accurate results compared to BIC on model-based Clustering (Estiri 2018).

After conducting the Clustering algorithm on the prepared data, interpreting the result of the cluster analysis is the most crucial phase. This will be more challenging when there are multidimensional clusters. Subject Matter Experts (SME's) should perform this interpretation. Distillation insight, the last stage of the process, tries to find those hidden correlations among datapoints, which are now formed into clusters.

Clustering programs based on student pass EFTSL

In institutional databases, one of the major levels of student data is the program level, in which each student/program has one record in a year. Student load refers to a measure that counts students in terms of full-time equivalence units in Australia, called EFTSL (Rouhi 2017) for higher education (HE) programs. The objective of this section focuses on the investigation of unknown patterns among university HE UGRD programs in 2020, based on the behavior of students on three aspects of the load. The three dimensions of student loads considered in this experiment are as follows:

- Certified_EFTSL; Total load that students acquired in the year,
- Pass_EFTSL; The portion of certified_EFTSL which successfully passed, and
- Cumulative pass_EFTSL; Total pass_EFTSL of the students from the starting of the

PROGRAM CODE	Enrolments Headcount	Total EFTSL	PASS EFTSL	CUMULATIVE PASS_EFTSL	PASS_EFTSL (avg)	CERTIFIED_EFTSL (avg)	CUMULATIVE PASS_EFTSL (avg)
Program1	545	391.39	320.85	625.76	0.59	0.72	1.15
Program2	48	34.04	30.92	50.78	0.64	0.71	1.06
Program3	566	394.44	322.56	580.22	0.57	0.70	1.03
Program4	123	102.76	90.00	134.92	0.73	0.84	1.10
Program5	325	263.17	251.26	457.86	0.77	0.81	1.41
Program6	258	184.47	163.19	308.13	0.63	0.72	1.19
Program7	206	153.39	119.23	225.30	0.58	0.74	1.09

Table 1: Structure of data for the first program experiment.

program

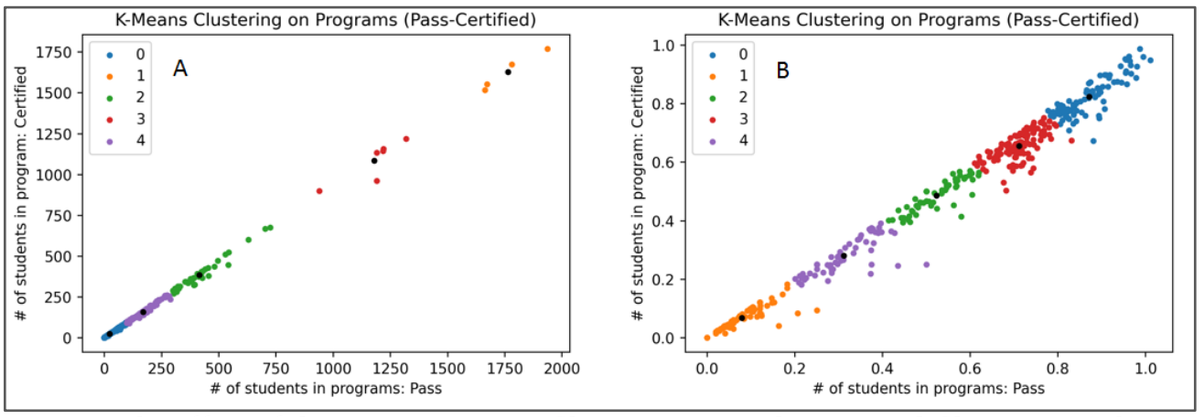

Figure 5: The effect of data preparation on Clustering. Two-dimensional Clustering with K=5 before averaging (A) and after averaging (B) on program pass_ and certified_EFTSL.

The value for pass_ and certified_EFTSL is maximum 1 in each year for a full-time student. Cumulative pass_EFTSL is considered in Clustering to investigate the possibility of correlation between pass load and the students in the program in the same the year; the more

years, naturally the higher the cumulative pass_EFTSL. The maximum value for a 4-year undergraduate program is 4.

The sample of input data for some programs is shown in Table 1. Enrolment headcount is added to the data to enable us to calculate the average figures for the above 3 types of EFTSL. This average calculation before feeding the data to a Clustering algorithm can be considered as a data preparation task. The raw values extracted from the database and the average values are shown on the left (Blue) and right (Green) columns in Table1.

To investigate the impact of averaging and raw values, Figure 5 depicts these in the form of a 2-dimensional Clustering on pass_ and certified_EFTSL. As can be seen clearly in this figure, the averaging forms the Clustering results more clearly. This is very similar to the impact of normalization shown in Figure 3.

In the next level of the experiment, averaging on raw data is considered; however, to investigate the correlation between pass_ and certified_EFTSL with the year of the program, the third dimension, cumulative pass_EFTSL, is added to the Clustering algorithm. The investigation on the 3 dimensions allows us to visualize the results on 3-D graphs; however, we should be aware that it is possible for Clustering to be applied on n dimensions. Also, n-D can be reduced to lower dimensions via the principal component analysis (PCA) technique, which is available in script languages like Python. Liang (2013) introduced the utilization of a distributed PCA in *K-means* Clustering. In this experiment, we have investigated the 9 Clustering sizes, with their BIC values and number of programs in each cluster bin shown in Figure 6. As can be observed, increasing the number of clusters reduces the BIC; however, it is our interpretation, our awareness of application constraints and our tacit experience that will finally lead us to select the most appropriate number of clusters. The results of the 3-D Clustering on 3,4 and 5 clusters are represented in Figure 6.

Insight extraction

Figure 6 illustrates the Clustering programs based on the behavior of students in passing and total EFTSL in different years of UGRD programs in one sample year (2020). The X axis represents pass_EFTSL and the Y axis represents the certified_EFTSL, with maximum

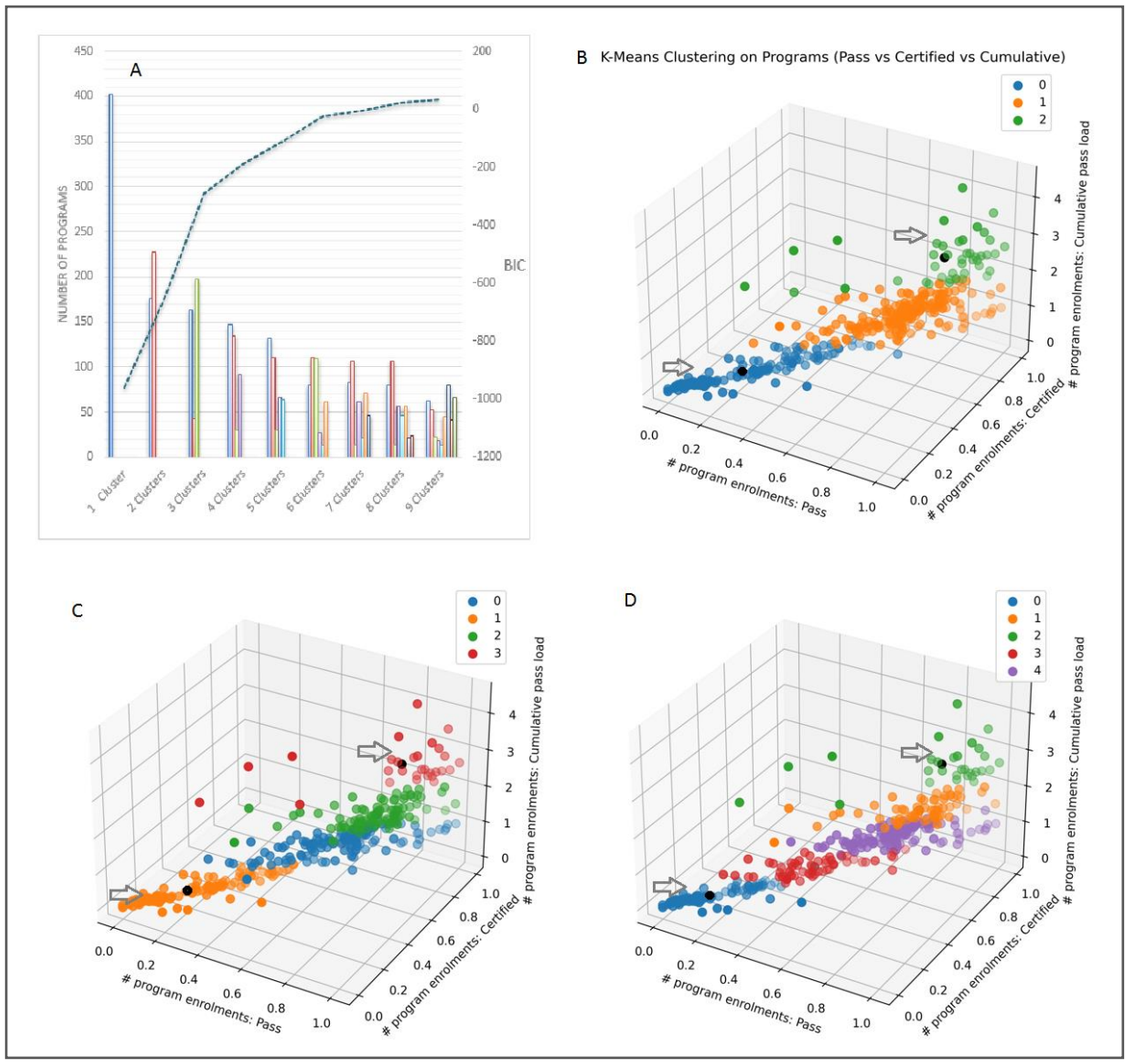

Figure 6: Investigation of the impact of different Ks in *K-means* Clustering and the BIC. A: Number of clusters from 1 to 9 and the BIC.
Results of number of K in 3-D Clustering on the 3 types of EFTSL, from 3 (B) to 5 (D) clusters.

values of 1 for a full-time student in each academic year. The Z axis represents the cumulative pass_EFTSL, and the maximum value for a 4-year UGRD program is 4. Nine Ks are investigated (from 1 to 9 clusters shown in Figure 6A) and the cluster formations of 3 of them are illustrated in Figure 6 (B, C and D for 3 ,4 and 5 clusters, respectively).

Extracting insight can be initiated by visual interpretation of the graphs. As can be seen in the Figure 6 – B, C and D, the propagation of the programs is shown by colored circles representing the clusters. The graphs clearly show two opposite groups of programs with their centroid points, which are as follows:

- The H_cluster, which includes programs with highest values in X, Y and Z axis; this cluster comprises cluster 2 in Figure 6_B, cluster 3 in Figure 6_C, and cluster 2 in Figure 6_D.
- The L_cluster, which includes programs with lowest values in X, Y and Z axis, this cluster comprises cluster 0 in Figure 6_B, cluster 1 in Figure 6_C, and cluster 0 in Figure 6_D.

The detailed results of the Clustering algorithm will provide us with a list of these two counter program clusters. The middle level clusters also contain valuable information. Sharing the results with SMEs and program managers would be useful to distill more insights not previously detected. Figure 7 illustrates how the average pass_, certified_ and cumulative pass_EFTSL of the H_ and L_clusters represent their aforementioned behavior with the magnitudes of their bar charts. The beauty of Clustering as an unsupervised machine learning algorithm is that it can clearly detect and group the UGRD programs based on their EFTSL load behavior and provide new and valuable insights for institutional researchers.

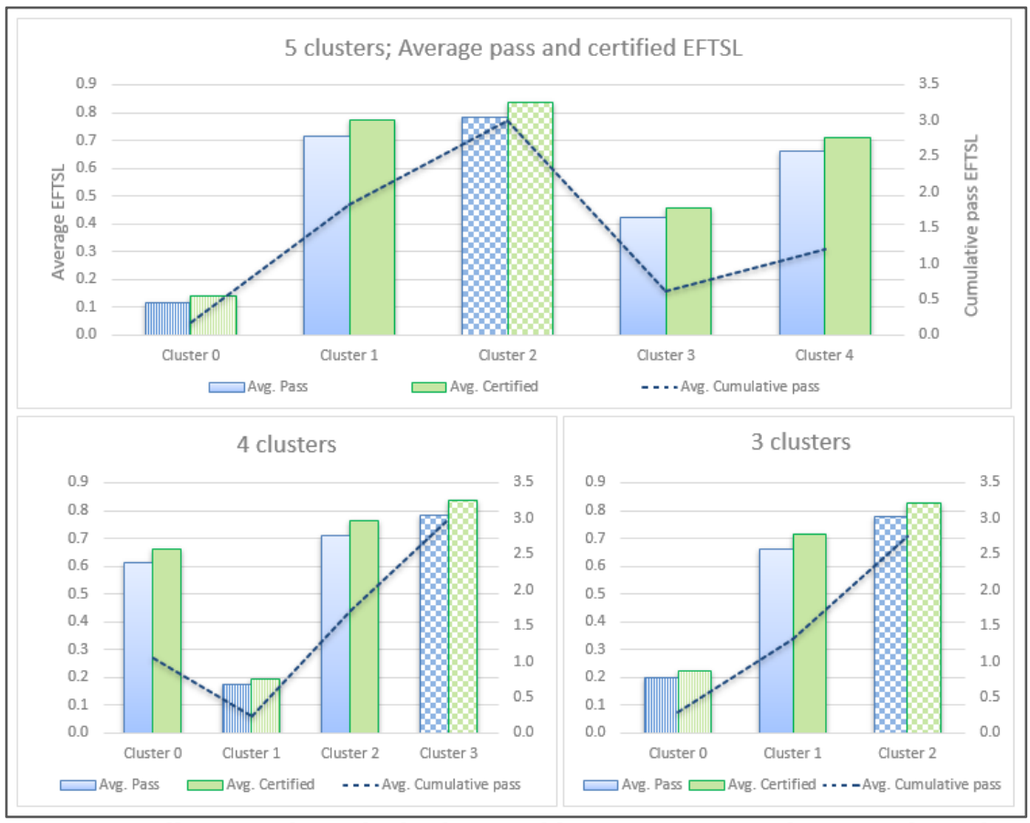

Figure 7: Bars illustrates how the highest and lowest average values are aligned with the Clustering results. The highest and lowest clusters are highlighted.

Conclusion

It is a fact that during the system design of institutional databases, the entities correlated to each other are detected and put together to form database tables. The normalization process in database design forces designers to avoid considering all the attributes in a flat single table, because this increases redundancy which is a red line in RDBMSs. Hence the result of the normalization process is the division of the data into correlated subgroups of data, a process which forms numerous tables in databases. However, it is possible to extract the different attributes from separated tables by applying joins on tables. It is well known that this cliché-structured data does not guarantee that all the possible correlations among the attributes (data columns) within the entities or among them (Tables) will be obvious or easily extractable utilizing conventionally designed database views and conventional structured data analysis.

With respect to the above-mentioned limitation and the unavoidable exponential growth of institutional data, utilizing Machine Learning (ML) algorithms is a bonus to overcome these barriers and to assist knowledge extraction and insight distillation. Unsupervised ML learning algorithms can analyze and cluster unlabeled datasets. these algorithms, such as Clustering, enables us to step further and go beyond the limitations of structured data. They are capable of automatically measuring the distances and grouping the datapoints into new clusters, without human interference. This process will help to detect hidden correlations among data, which will enable their grouping in a creative way.

The current research focuses on insight extraction based on the EFTSL (pass and certified load) patterns of students in the undergraduate programs in a given year. This research is just a sample of the Clustering techniques applied to student program data and resulting challenges. However, it can be applied on any level of institutional entities, such as course level data, human resources, equity groups, finances, etc. Finally, the 3 essential skills which enable us, when dealing with structured data, to perform the insight extraction process successfully are: accessibility to subject matter experts (SMEs) for extracting appropriate data; data preprocessing before applying Clustering algorithms; and, selecting the appropriate Clustering algorithms.

References

Huang, Z. *Extensions to the k-means algorithm for Clustering large data sets with categorical values*. Data mining and knowledge discovery, 2(3), pp.283-304. 1998

Kogan, J. Nicholas, C. & Teboulle, M., *Grouping Multidimensional Data Recent Advances in Clustering*, Springer Book, 2006.

Alzate, C. & Suykens, J.A., *Multiway spectral Clustering with out-of-sample extensions through weighted kernel PCA*. IEEE transactions on pattern analysis and machine intelligence, 32(2), pp.335-347, 2008.

Zhao, Q., Hautamaki, V. & Fränti, P., *Knee point detection in BIC for detecting the number of clusters.* In International conference on advanced concepts for intelligent vision systems (pp. 664-673). Springer, Berlin, Heidelberg, 2008.

Ayodele, T.O., Types of machine learning algorithms. *New advances in machine learning*, 3, pp.19-48, 2010.

Maimon, O. & Rokach, L., *Data Mining and Knowledge Discovery Handbook*, Springer Book, 2010.

Hastie, T., Tibshirani, R. & Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2011.

Liang, Y., Balcan, M.F. & Kanchanapally, V., *Distributed PCA and k-means Clustering.* In The Big Learning Workshop at NIPS , 2013.

Jordan, M.I. & Mitchell, T.M., *Machine learning: Trends, perspectives, and prospects.* Science, 349(6245), pp.255-260, 2015.

Rouhi, A. & Calderon, A., *Vector-based Models for Educational Institution Shape Analysis.* SEAAIR, 2017.

Potdar, K., Pardawala, T.S. & Pai, C.D. *A comparative study of categorical variable encoding techniques for neural network classifiers.* International journal of computer applications, *175*(4), pp.7-9. 2017

Estiri, H., Omran, B.A. & Murphy, S.N., *Kluster: an efficient scalable procedure for approximating the number of clusters in unsupervised learning*. Big data research, 13, pp.38-51, 2018.

Delua, J., *Supervised vs. Unsupervised Learning: What's the Difference?* IBM Analytics, Data Science/Machine Learning. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning. IBM Blog, 2021.

Scikit Learn Organization, https://scikit-learn.org/, Clustering examples

Acknowledgments

I thank Sean Lee from Analytics Centre of Excellence Department of RMIT University, for his considerable support, Nikhil Sobti and Basma Al-Mutawally, and all the Guild-day colleagues for providing guidance on deliberating the general concept of Clustering and its related techniques.