Data Science and Decision Making CE888 Assignment 1

Name: Muhammad Umar Farrukh Student ID: 1802565

***Abstract*—Goal of this project is to make use of machine learning techniques to perform different type of analyses on unsupervised data having no target value. Analyzing of data and performing of different cluster analysis techniques on categorical data is the main aim of this project. Feature selection by use of auto-encoder and transformation of data into particular categories will be made. Data-set selection and exploration of particular clusters is the primary aim of this project.Three different data-sets would be selected from different web-site and proper analysis would be performed making a detailed report on each of the categorization performed after analysis of data.**

I. Introduction

Main challenge of this project is to analyze the three different data sets and perform K-means clustering along with hierarchical clustering to categorize data sets into different clusters, the first data set which would be analyzed consists of Toyota company particular data which would allow to categorize the sale of car models along with their dominant features, second data set belongs to East-West airline in which travelling of passengers along different routes with their mile numbers is categorized, Third data set belongs to United States police. In third data set different categorical variables of police data defining the races involved in incidents in that particular regions will be categorized depending on the network of relation formed using auto-encoders and clustering.

II. Background

Unsupervised learning is a part of machine learning which allows us to identify and categorize data into particular different labels and clusters.Unsupervised learning allows us to explore the hidden features in a particular data set. Unsupervised learning prompts us to uncover previously unknown patterns in data. Data set used for unsupervised learning is unlabelled and is non-categorized.In unsupervised learning both the input and output variables are unknown.Unsupervised learning is usually deployed where separation of two different sets or clusters can be formed based on some inherent features provided in the data-set.Multiple different features describe the training network connection of features.It can be said that it is a self-organizing or adaptive learning algorithm as no future predicting features are provided to the network. Training features and input patterns are provided to the system which system arranges into form of categorizes or clustering. A group of training features are provided to the system at input layer, network associated weights are assigned over some kind of nodes of the output layer where the node with the highest value will be the successful candidate. It is mainly useful for clustering and association algorithm.[1]

A. Clustering

Clustering is a technique in data mining in which a group of data objects are taken as an input and output is a number of clusters obtained which are in form of groups called as clusters. Clustering is used mainly for pattern recognition, grouping data based on some particular instructions, classification. Number of mining techniques involve clustering like partitioning technique, hierarchical technique, gird-based technique, density-based technique, divide and conquer, random sampling.[2]
All these methods vary in terms of procedures executed for measuring similarity between clusters, construction of clusters by defining the threshold level, pattern of clustering that allow objects to belong cluster or another in terms of degrees and structure of algorithm. [3]

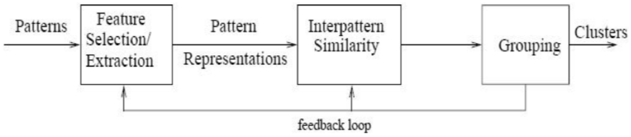


Fig. 1: Stages in Cluestering

[3][4] Typical clustering algorithm involves pattern representation which involves feature extraction or selection of features.After selection of features comes the pattern recognition of the data domain which is then followed by the clustering and if needed data abstraction and assessment of output is also done to get a higher accuracy result.Common working of the clustering is defined in the Figure2.

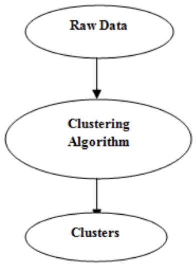


Fig. 2: Clustering

[3]

1) Hierarchical Clustering: Hierarchical clustering is a method that can be thought of as set of flat clustering method which is organized in a tree structure.Construction of clusters can be done recursively partitioning data set into top-down or bottom-up hierarchy.Among Hierarchical algorithms bottom-up tend to have more accuracy than the top-down approach but they have more computational cost. However the increase in computational complexity can not coincide with increased algorithmic complexity as the process of cluster hierarchy can be known as portioning operations.

Hierarchical clustering is divided into two different parts known as agglomerative clustering and divisive clustering.Agglomerative hierarchical based clustering defines the various distance measures such as single linkage, complete linkage.Agglomerative hierarchical clustering is also known as bottom-up approach in which the object initially represents the cluster of its own and clusters are merged iteratively until the desired cluster is obtained.Criteria defined in the algorithm are mainly minimum distance, maximum distance, average distance and center distance. [3]

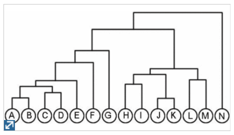


Fig. 3: Dendrogram for an agglomerative hierarchical clustering ,bottom-up approach

Single linkage produces long tail clusters whereas the complete tail clusters produce tight or compact clusters. According to the paper generalized linkage expression of the agglomerative clustering can be defined as: [5]

$$D(C_l,(C_i,C_j)) = \alpha_i D(C_l,C_i) + \alpha_j D(C_f,C_j) + \beta D(C_l,C_j)$$

$$+\gamma|D(C_l,C_i) - D(C_l,C_j)|$$

(1)

Top-down approach defines that objects belong to single root cluster and by use of iteratively partitioning defined clusters can be sub-divided into the required sub-clusters.

2) K-means Clustering: K-means clustering is a partitioning method which has a wide usage in scientific and industrial applications these days.K-mean discover clusters by finding k centeroids which are also defined as cluster representative or central points.K-means is applicable where only number of clusters are defined.It requires the of k in advance and performs in a local search procedure of finding the clusters.For a given k, algorithm is illustrated in four steps in figure 4. [6]

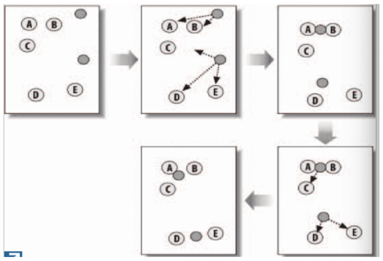


Fig. 4: K=2 illustration of K-means algorithm

3) Artificial Neural Network-Based Clustering: In terms of biology clusters perform the purpose of neurons scattered in the space.Every cluster is made with particular stimulated neuron and surrounding neurons.Neurons are arranged in particularly specified topology and than patterns are fed into the input layer of neural network.Self Organizing Map has the terminology of of producing from low dimension to high dimension. [7]

4) Fuzzy Based Clustering techniques: Fuzzy based clustering follows a traditional clustering approach in which that particular pattern belong to one cluster. Clusters are disjoint sets. Membership between different clusters is defined in the algorithm which has the probability to evaluate one pattern among all other defined patterns.Output of fuzzy clustering techniques are clusters but not a particular partition.Fuzzy clustering has a set of patterns which act as in fuzzy nature.It is similar to K-mean clustering except that a membership matrix is produced which defines the degree of membership to all other clusters formed. In a document of N numbers and clusters K than the beginning of membership matrix U is by N X K matrix. Element of this matrix represents the degree of membership.Now using this matrix U value of every fuzzy criterion is defined now. After execution of every function documents are than reassigned to clusters and to reduce the function value matrix U is recomputed.Process is repeated till all the values in U do not change.Pairs ordered in each clusters represents the document and membership value to cluster. [4]

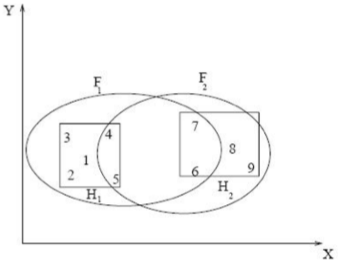


Fig. 5: Fuzzy Clustering

[8]

5) Search-based Clustering: Search-based clustering consists of more heuristic and smart approaches which are developed using optimization algorithm.Simulated Annealing, Deterministic Annealing and Particle Swarm Optimization are some examples of the search-based clustering algorithm [5]

B. Auto-Encoders

An auto-encoder is an unsupervised learning algorithm which applies methodology of back-propagation and sets the target values to be equivalent to input values.Auto-encoder learns the function and than tries to identify function so that output function is similar to input function.[9] Through identity function we can try to learn hidden facts about the provided unsupervised data. As an example if the input x are the pixel intensity values from a 100 pixels image.N = 100 and S_2= 50 hidden units, network is made to learn the compressed representation.In auto-encoders there is a neural network which reconstructs its output so an auto encoder

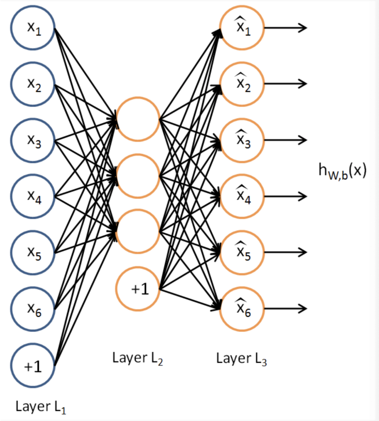


Fig. 6: Illustration of Auto-Encoder

is given an input of (1,0,0,0,1,0) the output formed by the encoder would be (1,0,0,0,1,0). The most important feature of the auto-encoder is in the hidden layer which has 5 dimensions and if 2 neurons are found in hidden layer than the encoder will receive 5 features and encode 2 features so it can reconstruct itself from the same input.So if a auto-encoder is trained with millions of data it will find weights which can minimize reconstruction error.

C. Activation Functions

In previous research studies Artificial Neural Network is made a base for data-modeling tool which acquires and presents complex input and output relations. Neural networks are alike human brain which gain knowledge through human brain and stores inner neuron connection strengths which are known as synaptic weights. Neuron first collects all its synapses after summing all influences acting on it. as the exciting influences become dominant, then the neuron fires and sends message to other through synapses.Neuron function model as simple threshold function. When neuron fires the combined signal strength exceeds threshold so in that case neuron value is provided by activation function.

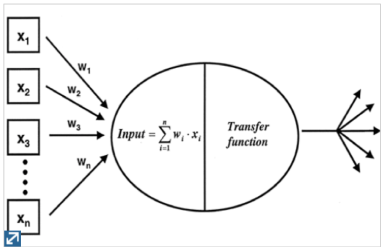


Fig. 7: Overview of acting neuron

[4] When input is applied on the weights and and they are further processed by the transfer function a net input value is generated depending upon the connection, if the total sum of weighted input is above a certain threshold then the neuron is fired by the activation function which is illustrated in the following diagram:

[10]

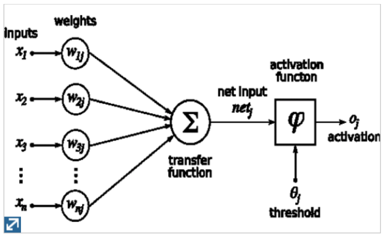
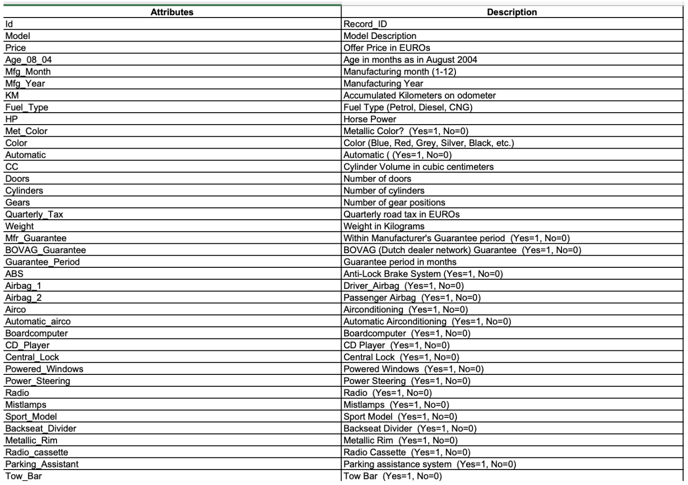


Fig. 8: Diagram of the inputs and their adaptive weights and finally comparing their outputs with the threshold function

III. Background

A. Data Collection

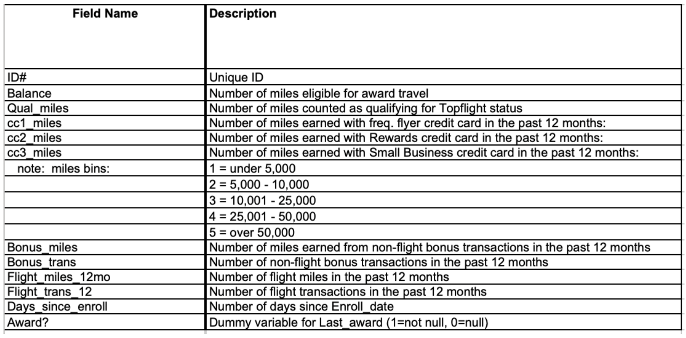
For development of research ideas on unsupervised learning many different sites were visited like Kaggle, UCL, git-hub and Data mining for business analytic's. Kaggle is one of the leading open-source website which is used by many students and analysts, most of the data-sets with their particular scripts and findings are uploaded on this website and with help of different people suggestions and comments the way of analyzing un-structured and structured data sets is made more interesting.Similarly git-hub and UCL provide free data sets with domain knowledge about the attributes involved in the data set. First data set was collected by surveying different sites and than concluding at a site named as Data Mining for business analytic's. It is a site with free domain knowledge site about many machine learning concepts and free data sets which are explained with some of the findings. Toyota company car data was exported from this site and the main goal to find from this data set was to explore the available attributes and provide the analysis that which model of that particular car has the highest sales. Similarly second data set was also extracted from this site which contains data related to East West airline and that data is used to analyze that among which route most of the passengers are travelling. Third data set is extracted from Kaggle website which is an open-source website. This data has most of the records related to the police arrests, main aim of this data set is to analyze the arrests and crimes of that particular region.



Attributes	Description
Id	Record_ID
Model	Model Description
Price	Offer Price in EUROs
Age_08_04	Age in months as in August 2004
Mfg_Month	Manufacturing month (1-12)
Mfg_Year	Manufacturing Year
KM	Accumulated Kilometers on odometer
Fuel_Type	Fuel Type (Petrol, Diesel, CNG)
HP	Horse Power
Met_Color	Metallic Color? (Yes=1, No=0)
Color	Color (Blue, Red, Grey, Silver, Black, etc.)
Automatic	Automatic ((Yes=1, No=0)
CC	Cylinder Volume in cubic centimeters
Doors	Number of doors
Cylinders	Number of cylinders
Gears	Number of gear positions
Quarterly_Tax	Quarterly road tax in EUROs
Weight	Weight in Kilograms
Mfr_Guarantee	Within Manufacturer's Guarantee period (Yes=1, No=0)
BOVAG_Guarantee	BOVAG (Dutch dealer network) Guarantee (Yes=1, No=0)
Guarantee_Period	Guarantee period in months
ABS	Anti-Lock Brake System (Yes=1, No=0)
Airbag_1	Driver_Airbag (Yes=1, No=0)
Airbag_2	Passenger Airbag (Yes=1, No=0)
Airco	Airconditioning (Yes=1, No=0)
Automatic_airco	Automatic Airconditioning (Yes=1, No=0)
Boardcomputer	Boardcomputer (Yes=1, No=0)
CD_Player	CD Player (Yes=1, No=0)
Central_Lock	Central Lock (Yes=1, No=0)
Powered_Windows	Powered Windows (Yes=1, No=0)
Power_Steering	Power Steering (Yes=1, No=0)
Radio	Radio (Yes=1, No=0)
Mistlamps	Mistlamps (Yes=1, No=0)
Sport_Model	Sport Model (Yes=1, No=0)
Backseat_Divider	Backseat Divider (Yes=1, No=0)
Metallic_Rim	Metallic Rim (Yes=1, No=0)
Radio_cassette	Radio Cassette (Yes=1, No=0)
Parking_Assistant	Parking assistance system (Yes=1, No=0)
Tow_Bar	Tow Bar (Yes=1, No=0)

Fig. 9: Diagram of the Toyota company data set attributes

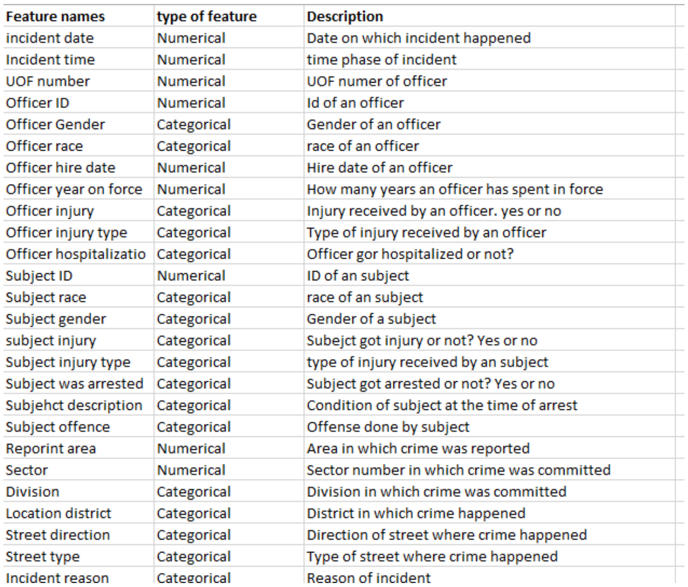
In Toyota company car data set particular different models of corolla car has been defined with different types of features like price, color, doors,cylinders, braking systems installed in that particular model, a total of 39 attributes with a total number of 1437 instances has been defined in this data set. Particular findings in this data set will revolve around the sale of that particular model with the manufacture year of the car.Most of the attribute are correlated with each other like the model of car is directly correlated with the manufacturing year, similarly color of car would be correlated with the price attribute and model of the car. Price range would be correlated with the model and count of features. All the attribute of this Toyota car company data set are defined in Figure 9 which is as follows:



Field Name	Description
ID#	Unique ID
Balance	Number of miles eligible for award travel
Qual_miles	Number of miles counted as qualifying for Topflight status
cc1_miles	Number of miles earned with freq. flyer credit card in the past 12 months:
cc2_miles	Number of miles earned with Rewards credit card in the past 12 months:
cc3_miles	Number of miles earned with Small Business credit card in the past 12 months:
note: miles bins:	1 = under 5,000
	2 = 5,000 - 10,000
	3 = 10,001 - 25,000
	4 = 25,001 - 50,000
	5 = over 50,000
Bonus_miles	Number of miles earned from non-flight bonus transactions in the past 12 months
Bonus_trans	Number of non-flight bonus transactions in the past 12 months
Flight_miles_12mo	Number of flight miles in the past 12 months
Flight_trans_12	Number of flight transactions in the past 12 months
Days_since_enroll	Number of days since Enroll_date
Award?	Dummy variable for Last_award (1=not null, 0=null)

Fig. 10: Diagram of the East West airline company data set attributes

In East-West airline a total number of 12 attributes are defined with a total number of 4000 instances. Main goal of this data set is to learn more about its customers by studying the different flying patterns they are using to travel from one particular destination to another destination.All the particular attributes of the data set along with the description are explained in the figure 10.



Feature names	type of feature	Description
incident date	Numerical	Date on which incident happened
Incident time	Numerical	time phase of incident
UOF number	Numerical	UOF numer of officer
Officer ID	Numerical	Id of an officer
Officer Gender	Categorical	Gender of an officer
Officer race	Categorical	race of an officer
Officer hire date	Numerical	Hire date of an officer
Officer year on force	Numerical	How many years an officer has spent in force
Officer injury	Categorical	Injury received by an officer. yes or no
Officer injury type	Categorical	Type of injury received by an officer
Officer hospitalizatio	Categorical	Officer gor hospitalized or not?
Subject ID	Numerical	ID of an subject
Subject race	Categorical	race of an subject
Subject gender	Categorical	Gender of a subject
subject injury	Categorical	Subejct got injury or not? Yes or no
Subject injury type	Categorical	type of injury received by an subject
Subject was arrested	Categorical	Subject got arrested or not? Yes or no
Subjehct description	Categorical	Condition of subject at the time of arrest
Subject offence	Categorical	Offense done by subject
Reporint area	Numerical	Area in which crime was reported
Sector	Numerical	Sector number in which crime was committed
Division	Categorical	Division in which crime was committed
Location district	Categorical	District in which crime happened
Street direction	Categorical	Direction of street where crime happened
Street type	Categorical	Type of street where crime happened
Incident reason	Categorical	Reason of incident

Fig. 11: Diagram of the Dallas Texas USA data set attributes

Dallas Texas has a total attributes of 29 with a total number of estimated instances 25000. Main goal of this data set is to cluster the different kind of crime which are happening in Dallas Texas region and the particular use of force used by police in that particular region. Different types of races both for police officers and the suspects will also be determined.In this data set most of the attributes are correlated with each other. Subject race is directly correlated with the officer race, similarly the officer race will be directly correlated with the age factor, total number of crimes would be directly correlated with the location attribute.

IV. Experimental Analysis

For experimental analysis the main tool usage will be Python with Jupyter notebook acting as Development enviorment. Different machine learning libraries will be used like Sci-kit, Scipy, Numpy, Pandas, Bootstrap, and tensor flow library for deep learning analysis. First all the data sets would be normalized using different normalization techniques such as calculating mean of that particular row or column if that value is appearing as null. After the cleaning of data total counts of particular attributes would be found for analyzing the categorical values. Having a good insight analysis of the data set and the attributes will lead us to apply different types of clustering techniques. For clustering techniques first hierarchical clustering will be applied which will be giving results in form different clusters, using hierarchical clustering particularly the top-down or bottom-up approach will be followed depending upon the situation, than dendograms will be made which will give us an insight evaluation of the clusters performed at a particular stage. After getting an insight analysis of the different clusters performed at different levels we will perform K mean clustering which will be giving the categorical analysis of the data set. Getting categorical analysis will lead us to the use of auto-encoders which will find the hidden structure in the data set and once the hidden layer is found out particular categorical clusters will made. Auto-encoders will be following the iterative process in which after forming a particular cluster it will form another inside it self until no particular cluster can be made. For analysis purpose grid search will be used and which will be giving the optimal search values for the clustering techniques applied. Optimal search results will be defining the features and from this we can do feature selection which can increase the result of that particular clustering technique. Other than grid search technique cross-validation technique will also be applied which can be also used for training the particular algorithm according to the given features. For analyzing the final result completeness score will be recognized for a given particular algorithm. [11] Three different sections will be formed and the assignment of input layer to output layer generated results will be analyzed that which particular clustering technique has the highest evaluation score. transparency of the input layer with the output layer will be determined.

V. Conclusion

In this experiment we will apply three different techniques Hierarchical clustering with top-down or bottom-up approach, K-means clustering and auto-encoders after forming clusters, cross-validation tables and grid-search will be applied to analyze the results of clusters and particular features. After that feature selection will be performed to increase the accuracy rate which will be measured using completeness score technique.

References

[1] H. U. Dike, Y. Zhou, K. K. Deveerasetty, and Q. Wu, "Unsupervised learning based on artificial neural network: A review," in *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*. IEEE, 2018, pp. 322–327.

[2] P. J. Kaur *et al.*, "A survey of clustering techniques and algorithms," in *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2015, pp. 304–307.

[3] A. K. Mann and N. Kaur, "Review paper on clustering techniques," *Global Journal of Computer Science and Technology*, 2013.

[4] S. Mythili and E. Madhiya, "An analysis on clustering algorithms in data mining," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 1, pp. 334–340, 2014.

[5] B. F. Momin and G. B. Rupnar, "Keyframe extraction in surveillance video using correlation," in *2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*. IEEE, 2016, pp. 276–280.

[6] Z. Nazari and D. Kang, "A new hierarchical clustering algorithm with intersection points," in *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. IEEE, 2018, pp. 1–5.

[7] Y. Bengio and O. Delalleau, "On the expressive power of deep architectures," in *International Conference on Algorithmic Learning Theory*. Springer, 2011, pp. 18–36.

[8] R. Agrawal and M. Phatak, "A novel algorithm for automatic document clustering," in *2013 3rd IEEE International Advance Computing Conference (IACC)*. IEEE, 2013, pp. 877–882.

[9] Y. Qiu, W. Zhou, N. Yu, and P. Du, "Denoising sparse autoencoder-based ictal eeg classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 9, pp. 1717–1726, 2018.

[10] S. P. Singh and V. M. Srivastava, "Implementation of digital neuron cell using 8-bit activation function," in *2011 Nirma University International Conference on Engineering*. IEEE, 2011, pp. 1–4.

[11] N. Guntamukkala, R. Dara, and G. Grewal, "A machine-learning based approach for measuring the completeness of online privacy policies," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015, pp. 289–294.