CE888 Data Science and Decision-Making Assignment 1

Submitted as part of the requirements for:

CE888 Data Science and Decision Making

Name: Julien Gergi Sarkis

Tutor: Matran-Fernandez, Ana

Date: 21 February 2019

Contents

Abstract

This age we are living in is depending on the big data, nearly everything we do daily is being transformed by the big data like engineering, medicine, business and even social life. Analysing the data is the challenge which require many algorithms; this is the field of data science, in this report we will be working on the unsupervised learning algorithm clustering, we will use neural networks to learn favourable features for clustering which will happen by training autoencoders.

Introduction:

Data science is a field that uses algorithms to get information from data and it consists of a number of phases, first the collection of the data which is one of the hardest part of data science, after the data have been collected it has to be processed, after that exploration and visualization of the data is helpful to get an insight on the data and the features of it, then analysing or applying machine learning techniques to the data and after all of that deciding based on all the steps that were done beforehand.

The purpose of this report is to learn good features for clustering using the method of training auto-encoders, there are various tasks to be done firstly we will be collecting data from UCB archive, specifically three reasonably sized datasets we willl have to load and inspect them, after that clustering of all the dataset will be done using Scikit-learn clustering then a training of an auto-encoder for the data that will be used as inputs for the clustering algorithms. The next task will be training the same auto-encoders but changing the middle layer using softmax features, then we will use the layer with the biggest value for the assignment and last, evaluation of the method used will be done.

The reason that data science is important is that it analyses data collected from a certain topic in a certain time and it can help predict how to work in the future or how to change specific things related to the topic which could help making a more successful outcome, basically optimizing the work that has to be done based on data from previous work in the same domain.

Literature Review:

According to [1], the era we are living in which is the big data era, there are big numbers of various valuable data of different credibility that is collected easily or achieved with high speed, and in these data, there are valuable information for the data scientists to discover and analyse. One the famous unsupervised learning methods is clustering, and it is used to arrange a set of unlabelled data. In [1], they also went on and categorized clustering algorithms such as partitioning approaches, density-based approaches, hierarchical approaches, grid-based approaches, distribution approaches and other approaches.

An attention in the recent times on the big data shows that the companies are aware of the potential of data collected by the information systems of nowadays, and most researches are covering the investigation of analytical intelligence that can be used for many useful predictive tasks [2]. In [3], Clustering was explained as an unsupervised method to group objects that are close to each other, it has important aspects such as feature selection, distance metrics and many grouping methods such as K-means which is one the most used clustering algorithm, also deep clustering techniques were discussed that they used auto-encoders to learn low dimensional data representations; The auto-encoders learn how to keep the most important features for the distribution of the data. However, in [4] they also wrote about the goal of clustering which is the grouping of similar objects and that the importance of clustering is increasing due to new areas of use for example data mining.

According to [5], many big data applications needs big number of data with high dimensions, which requires an efficient process, which causes a challenge in the high dimensional data those result the outliers. However, in that paper, unified unsupervised Gaussian mixture auto-encoder to detect outlier in the high dimension data.

Methodology:

In this experiment, learning good features for the unsupervised clustering method by training auto-encoders is the expected result from the analysis. The methods that are going to be used in the experiment is K-means clustering using the Scikit-learn library in python, training of auto-encoders to know which are the most important features to be used as inputs in the clustering will also be used, and then evaluation of the methods on many cluster sizes will be done in the end.

Three data sets were collected from archive datasets:

1. Human activity recognition using smartphones data set which is from recording 30 subjects performing daily activities while putting on a waist smartphone with embedded inertial sensor. The characteristics of the dataset are multivariate and time-series, the tasks are classification and clustering, it has 10299 instances with 561 attributes and no missing values. The daily activities that were done involves walking upstairs and downstairs, sitting, standing and laying

2. Gesture phase segmentation data set which is made by features gotten from seven videos of people gesticulating. The characteristics of the dataset are multivariate, sequential and time-series, the tasks are classification and clustering, it has 9900 instances and 50 attributes with no missing values. Each video has two files a raw file containing the position of the hands, wrists, head and spine of the person and a processed file with the velocity and acceleration of the hands and wrists.
3. Grammatical facial expressions data set which has the characteristics of multivariate and sequential, real attributes characteristics, classification and clustering tasks. It has 27965 instances and 100 attributes with no missing values. This data set is made from eighteen videos using Microsoft Kinect sensor, the user will five times in front of the sensor make five sentences in sign language that needs facial expressions.

Experiments:

Using Python language, many python libraries and PyCharm software, we will first evaluate and visualize the data set so we know which methods to use. Auto-encoders will be built and trained on the data and will be used as input features for the clustering. K-means algorithm will be used to cluster data by separating samples in n groups of equal variances.

Discussion and Evaluation:

Two metrics will be used for evaluations which are completeness and silhouette coefficient, completeness is assigning all of the data points that are members of a single class to a single cluster, and the silhouette coefficient indicates +1 if the sample is far away from the neighbouring cluster, 0 if the sample is very close to the decision boundary between two clusters and negative values will indicate that the samples might have been assigned to the wrong cluster.

After the evaluation of the algorithms used in this experiment, improvement will be done by repeating the steps with different parameters or adding a technique to the experiment depending on what will be needed after the evaluation and the visualization of the data.

Conclusion:

Working on data sets using unsupervised learning techniques requires following number of steps which will be used in this experiment starting with visualizing and analysing the data to evaluation of the method used and repeating all the steps until Achieving the best results possible.

References:

[1] Dierckens, K., Harrison, A., Leung, C. and Pind, A. (2017). A Data Science and Engineering Solution for Fast K-Means Clustering of Big Data. *2017 IEEE Trustcom/BigDataSE/ICESS*.

[2] F. Rocha Silva, "Analytical Intelligence in Processes: Data Science for Business", *IEEE Latin America Transactions*, vol. 16, no. 8, pp. 2240-2247, 2018. Available: 10.1109/tla.2018.8528241 [Accessed 20 February 2019].

[3] P. Dahal, "Learning Embedding Space for Clustering From Deep Representations", *2018 IEEE International Conference on Big Data (Big Data)*, 2018. Available: 10.1109/bigdata.2018.8622629 [Accessed 20 February 2019].

[4] C. Fraley and A. Raftery, "Model-Based Clustering, Discriminant Analysis, and Density Estimation", *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611-631, 2002. Available: 10.1198/016214502760047131.

[5] W. Liao, Y. Guo, X. Chen and P. Li, "A Unified Unsupervised Gaussian Mixture Variational Autoencoder for High Dimensional Outlier Detection", *2018 IEEE International Conference on Big Data (Big Data)*, 2018. Available: 10.1109/bigdata.2018.8622120 [Accessed 20 February 2019].

[6]https://scikitlearn.org/stable/modules/generated/sklearn.metrics.completeness_score.html#sklearn.metrics.completeness_score

[7] https://scikit-learn.org/stable/modules/clustering.html

[8] https://blog.keras.io/building-autoencoders-in-keras.html

[9] Xie, Junyuan, Ross Girshick, and Ali Farhadi. “Unsupervised deep embedding for clustering analysis.” International conference on machine learning. 2016.

[10] https://archive.ics.uci.edu/ml/datasets.html

Plan:

1. Selecting and inspecting three different datasets from archive website. (1 week)
2. Using K-means clustering technique, all the data sets will be clustered. (2 weeks)
3. Building auto-encoders to train on the data so that it will select the best features as inputs to clustering. (2 weeks)
4. Evaluation of the clustering and repeating until reaching the best results. (1 week)