ISTANBUL TECHNICAL UNIVERSITY

Plot Based Clustering Using Convolutional Autoencoder Representations

1. INTRODUCTION

In the contemporary world, the machine learning and deep learning applications are becoming more and more beneficial for many industries with the increased availability of all kinds of data in any form such as text, audio, image and tabular. Marketing is one of the fields that these applications have started been widely used especially in the last decade or so. With the advancement in communications technologies, measuring and tracking consumer behaviour have started to become an important battleground for companies that are looking to maximize their profits. Applications of these technologies using consumer behaviour data became an increasingly interesting topic, widely named as MarTech as stated by Hauser (2007).

Customer segmentation has always been an essential topic of marketing for a long time. Customer segmentation is usually been made with different types of data like. While Demographic segmentation includes features like Age, Marital Status, Job, Gender, Income, Occupation; Psychographic Segmentation included features like Interests, Activities, Opinions, Values. The earlier applications of customer segmentation were mainly based on non-behavioural features of customers such as demographics and socioeconomic status. The third and most popular type of segmentation which is behavioural is getting more attraction with the access to behavioural real-time shopping data like Usage, User Status, Occasions and Brand selection as shown by Best (2004).

However, behavioural data of customers such as purchase, or usage records of a product are becoming more and more available for many companies in the recent years. Consequently, analysing and segmenting the customers based on their behavioural data is a common concern for all companies in various sectors. Companies pursue the goal of creating more efficient and beneficial ways of communication with their customers by utilizing their knowledge of customer segments. Therefore, the segments obtained by the segmentation analysis should be partitioned well and give interpretable, generalizable and useful insights about the group of customers they contain. As in many other tasks, machine learning techniques are commonly utilized to create the ambitioned customer segmentation models. Because of the nature of the segmentation task, clustering, which is a very commonly used unsupervised machine learning method, algorithms are mostly applied to create segmentation models. As a result, the chosen clustering algorithm, data representation which is fed into the algorithm, and dissimilarity metrics are the most important factors while creating a customer segmentation model using clustering techniques.

In this study, the aim is to create a novel and efficient approach of representing a tabular dataset to the clustering algorithms with the purpose of creating well separated and interpretable customer segments. The proposed idea is basically composed of two steps; converting the dataset to image form by creating plots for each individual customer based on the tabular dataset, and training convolutional autoencoders on the image dataset to obtain the latent space vectors of each image, while applying clustering algorithm on the obtained latent space vectors which can also be referred as data representations. The scientific intuition and motivation for this attempt has been explained and discussed in the rest of the study by utilizing various ideas and applications in the literature.

The rest of the study is organized as follows: the explanation of the key concepts of clustering and representation learning using autoencoders, analysis of similar applications in the literature, model building and analysing results and, finally the conclusion part to summarize the study.

2. KEY CONCEPTS

In this part, the key concepts of clustering and deep learning have been related to the proposed approach.

2.1. Clustering

2.1.1. Clustering Algorithms

Clustering algorithms are the most commonly applied and popular algorithms in unsupervised machine learning branch. The aim of clustering algorithms is mainly creating a partitioning of a dataset which keeps the similar observations in the same subgroup while keeping the different observations in separate subgroups. There are many clustering algorithms such as K-Means, hierarchical clustering, and partitioning around medoids which have different advantages and disadvantages by the means of computational cost and the capabilities. The following explanations of the algorithms are summarized versions of the relevant chapters of Hastie et al. (2017).

K-Means algorithm is basically a simple clustering approach to dive a data set into K distinct groups. The main idea of the K-means algorithm is to minimize the total within-cluster variation, which means total difference between observations. The distances between observations are calculated by using Euclidian distance, and it is a computationally cheap algorithm.

Hierarchical clustering is a clustering technique which takes a dissimilarity matrix and a linkage criterion as inputs and gives clusters as an output. The dissimilarity matrix can be formed by using any distance metric in this algorithm. It creates an advantage to the algorithm by the means of enabling the researches to create clusters with different points of views using the same dataset.

Partitioning around medoids (k-medoids) is a divisive clustering algorithm which is highly similar to k-means clustering method. However, in k-medoids the cluster centers are assigned to one of the observations while in k-means it is not necessary. Another difference between is that arbitrary distance metrics such as optimal matching event distance, correlation distance and hamming distance can be used with partitioning around medoids technique, while the k-means algorithm uses Euclidian distance.

2.1.2. Distance Metrics

As briefly explained, the clustering algorithms basically finds the distances between observations and group the observations by comparing their pairwise differences.

Consequently, the selected distance metric is a very fundamental hyperparameter to obtain the clusters which are separated as pleased. In this aspect, the performances of clustering algorithms which use pairwise distance matrices as input, such as hierarchical clustering, are highly dependent on the preference of distance metric.

For example, correlation-based distance metrics, such Pearson correlation distance, are used to create clustering depending on the correlation between the attributes of samples. Consequently, while two observations can be found very far from each other using Euclidian distance, they can be found very close using Pearson correlation distance.

This situation is a big motivation for this study, because it shows that even if the dataset and the algorithm is constant, the chosen way of understanding the changes all the results. In the lights of this information, it can be said that creating plots of samples to analyse them with from different angles of look may create unique and beneficial solutions.

2.2. Representation Learning

Representation learning, also referred as feature learning, is a common concern for artificial intelligence and data science researchers. As stated by Bengio et al. (2013), the goal of progressing towards and creating artificial intelligence can only be achieved by making it learn to detect and use the hidden beneficial information from the raw visual or any other form of data to understand the surrounding environment. Deep learning algorithms are outperforming the other machine learning algorithms by learning the data representation themselves in the tasks such as speech recognition and signal processing, object recognition, natural language processing. However, when the data is unlabelled, there is a natural need for different kinds of feature learning algorithms. As asserted by Min et al. (2018), in unsupervised clustering applications, high dimensional data such as image should be represented by a lower dimensional feature space to create more efficient models. One of the most widely used algorithms to learn representations of unlabelled data is autoencoders.

Autoencoders are deep learning architectures which learns representations from raw data by using it directly as both input and output.

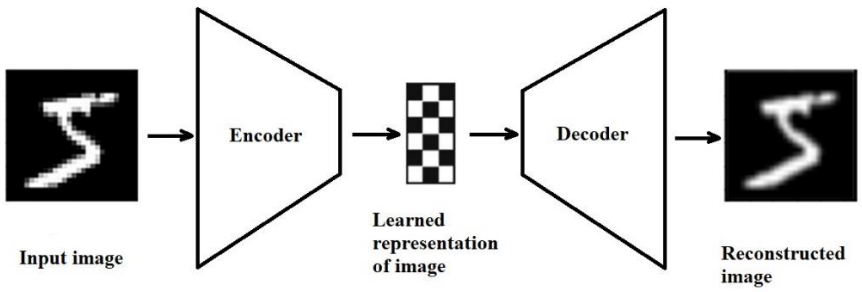


Figure 1. An illustration of a basic auto encoder which learns the representation of a sample image from MNIST handwritten digit dataset.

Basically, an autoencoder is formed by two parts of stacked hidden layers which are called encoder and decoder as Figure 1. shows. Bengio et al. (2013) shows that the encoder function

maps the input (raw data) to a feature (latent) space, and the decoder function maps the feature space back to the input space. The loss function is called reconstruction error which calculated as the difference between the input and output data. The ultimate goal of the algorithm is to encode the image to a lower dimensional representation which can be decoded to the original image by minimizing the reconstruction loss.

3. Related Studies:

3.1. Image Clustering Applications

Image clustering is a popular machine learning task and is studied by many researches in the literature. As proposed by Min et al. (2018), many of the studies in the literature are applying feature extraction and clustering operation as different steps. However, there are many studies which are proposing to combine the clustering and feature extraction processes to create more efficient algorithms. As shown in the study of Min et al. (2018), there many alternative models created in the literature to combine clustering and with representation learning algorithms which are utilizing different kinds of feature learning algorithms such as autoencoders, convolutional neural networks, variational autoencoders and generative adversarial networks. Caron et al. (2018) asserts that the generative model based feature learning algorithms have a less efficient performance in clustering domain. Moreover, in Min et al. (2018) it is shown that the generative model based architectures of deep clustering are much more computationally expensive. On the other hand, the convolutional neural network based architectures are not guaranteed to learn a decent feature space because this kind of architectures do not use a loss function related with the reconstruction error as stated in Min et al. (2018). Consequently, among many alternatives in the literature, autoencoder based models are seemingly the most promising models to successfully cluster images by their ability to learn consistent and representative feature space for images and applicability with many alternative clustering methods.

3.2. From Tabular to Image Transformation of Data

Image data is analysable and able to be clustered efficiently by using representation learning techniques. However, the main intuition behind this study suggests using plots (as image input) of numerical tabular data to achieve better clustering performance. Although there are not many studies utilizing and trying this method in the literature, there are some studies showing the potential of this approach. In the literature there are many applications of converting tabular data to image to create machine learning models in the finance discipline. The main focus of those studies is generally stock price classification or prediction to help decision making process of buying or selling stocks.

In the study of Kim and Kim (2019), it is stated that the visualization technique applied in the transforming process both beneficial for cleaning the noise from tabular data, and important to increase the performance of the model. From this perspective, they created many alternative plotting techniques on their data, which are called candle-stick chart, line chart, F-line chart, and bar chart to find the best representation. It is important here to emphasize that all the charts

are created from the same tabular input and representing the same information in various visual ways. On the other hand, Kim and Kim (2019) states that fusing charts is a good practise to improve the representation of data and showed that it increases the accuracy of their forecast significantly comparing to usage of single charts.

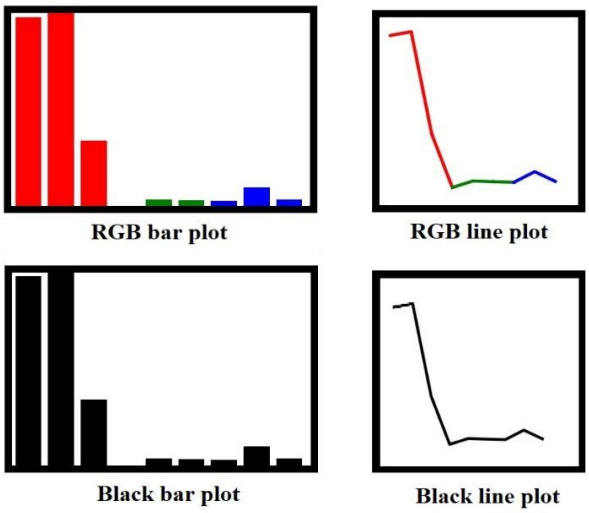


Figure 2. Example plot styles of a sample from the dataset which is aimed to be clustered. Basically, there are RGB and black coloured bar and line plots of the same observation (tabular input).

In Figure 2., a sample from the dataset is illustrated as different kinds of plots. As can be seen, although the same tabular input is used, the obtained image data shows quite different features. This situation is highly promising to create more interpretable clusters by the visual representation of choice. Another highly important aspect of creating visualizations of tabular data to learn different representation spaces is the colouring. Unlu (2020) suggested a novel visualization technique of candlestick charts by colouring certain parts of the charts with red, green and blue to create more analysable kind of image data for autoencoders. Because of the numbers channel are increasing in RGB images, the image contains more information when it is coloured. Consequently, this approach may enable researches to implement some feature engineering ideas, which are not applicable on tabular raw datasets, by changing the colours of their plots.

Another example study which uses from tabular to image transformation of the stock price data is Chen and Tsai (2020). The difference is that while Kim and Kim (2019) aims to make forecast analysis, Chen and Tsai (2020) is proposing a classification approach. In the research, it is stated that the transformation of candle stick charts to gramian angular fields representation, which is an image representation for time series data, has increased the model accuracy by preventing underfitting. In conclusion, it can be said that the usage of different kinds of image representations causes high changes in the performance of the classifier.

Although in financial environment this approach is used for supervised classification and/or regression tasks related to stock price forecasting for decision making, the study of Ryu et al. (2019) apply the approach of converting tabular data to images (by plotting) on an unsupervised clustering task in customer segmentation domain, which is almost perfectly parallel with the suggested idea in this study. Their study is aiming to cluster household by the yearly electricity

load consumption data. To achieve this, they used the input data in both tabular and image form to create different features of households from both types of data. The plotting type is a kind of heatmap in RGB form which preserves and resembles the structure of the time series data of yearly load profiles. The resulting clusters shown in the paper are promising for future studies, because of their distinguishable visual representations.

4. Model Building

4.1. Data Preparation

In this study, a dataset which is formed by the FMCG spendings of households in a certain period is used. The spendings are at first categorized into 3 main FMCG categories as food, cleaning and personal care products. Then each individual main category is divided into 3 price segments as premium, value-for-money and economy products. To normalize the data, instead of using the spending value, the percentages of price segments at each category are used as the input data. The dataset is created synthetically.

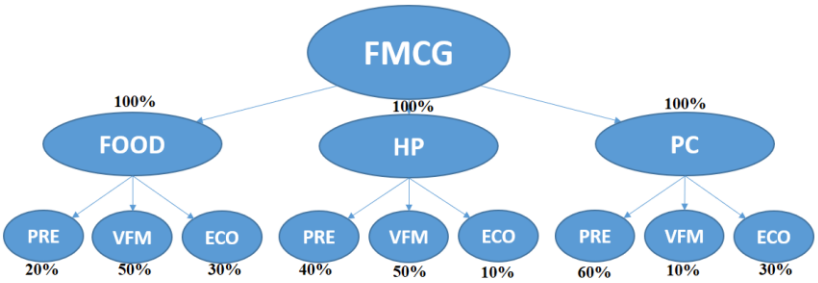
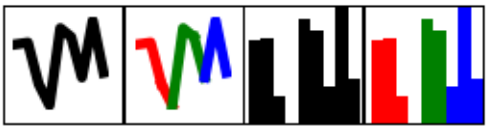


Figure 3. The steps of data preparation of an arbitrary household.

In the final form of the data, there are 9 different attributes for each of 14293 households. On Figure 3., an example household data is shown to illustrate the final form of the tabular dataset.

In the second part of data preparation, to test the idea of representing the same data with different styles would result in different clusters, 4 different plots for each row (resembling a household) are created. These plots are 64x64 sized line and bar plots in grayscale (with one channel) and RGB (with three channels) forms.



FOOD_PRE	FOOD_VFM	FOOD_ECO	CLEANO_PRE	CLEANO_VFM	CLEANO_ECO	PC_PRE	PC_VFM	PC_ECO
1441681	1478078	505875.5	0	61395.16	54419.75	49421.54	149790.9	57751.6

Figure 4. Different representations of the same sample of data. Above there are image representations and below the raw tabular form.

As the result 5 different datasets, which are different representations of the same data, including the tabular, grayscale and RGB line plots, and grayscale and RGB bar plots are generated to

create clustering models from each of them individually. In Figure 4., an example household's different representations are shown.

4.2. Model Architecture

The proposed model in this study is basically a convolutional autoencoder which is trained using a combination of the reconstruction loss and a custom-built clustering loss. The main intuition behind adding the clustering loss to training is to force the autoencoder model to create latent representations which performs well at the clustering task as well.

$$\min_{\theta}\left(\sum_{i=1}^{N}\left(\frac{(x_i-\hat{x}_i)^2}{N}\right)+\lambda\sum_{i=1}^{N}\left(\frac{(\hat{x}_i-m_i)^2}{N}\right)\right) \tag{1}$$

In Eq. 1., the loss function used to train the convolutional autoencoder model. While the variable $\hat{x}_i = \mathrm{f}(\mathrm{g}(x))$ shows the corresponding reconstructed image, m_i is the corresponding reconstructed version of the medoid of the cluster which x_i is assigned to. By creating this loss function, it is aimed to find latent variables which are both preserving the important information in the image data and resulting in visually uniform clusters by optimizing the neural network parameters using gradient descent algorithm.

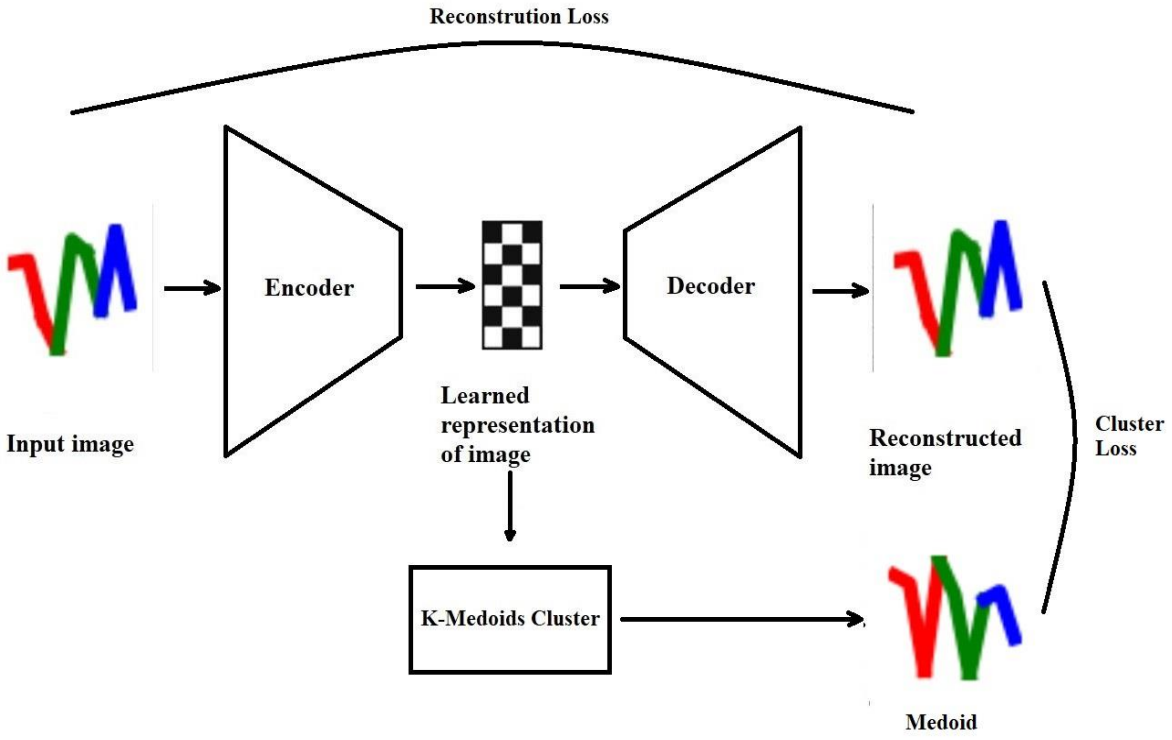


Figure 5. A simple diagram of the proposed model.

Figure 5., shows the model architecture. Basically, each batch of dataset is at first past through the encoding layers to create latent representations. The latent representations are at first used to create clusters with partition around medoids algorithm, and then past through the decoding layers to create reconstructed images. For each batch, the combination of reconstruction loss

and clustering loss, which is shown in Eq. 1., is calculated to update the parameters of all the network.

5. Results

After training models with all the different representations of the data, the resulting clusters and their performances are compared. To compare the models numerically, the custom-built clustering loss has been calculated for each clustering model.

Table 1. The comparison of clustering performances of the models using custom-built cluster loss. The rows are different models trained with different datasets, and the columns are their performances on image datasets.

	BWline Data:	**RGBLine Data:**	**BWBar Data:**	**RGBBar Data:**
BWLine_loss:	**0.0936**	**0.0673**	0.1233	0.0859
RGBLine_loss:	0.1031	0.0743	0.1387	0.0972
BWBar_loss:	0.1002	0.0716	**0.1167**	**0.0813**
RGBBar_loss:	0.1004	0.0719	0.1201	0.0835
Base_loss:	0.1011	0.0723	0.1190	0.0831

As Table 1. shows, the autoencoder based clustering models improved the performance of the base model, which is built by the tabular data. However, there is no model with the best scores for each type of data. Consequently, it cannot be said that any model is the best to represent the dataset. As the result, it can be concluded that the proposed method may improve the clustering model by means of being visually uniform, however the performance is highly dependent on the evaluation criteria and the chosen representation of the data.

On the other hand, the expectation before experiments was to have better clusters using the RGB image datasets, because the colouring is made category wise to

6. Conclusion

As the result of this report, it can be said that using visual representations of tabular data is an application which became available for researchers recently with the improvements in the deep learning and computer vision related fields. Even though there are not many applications in the literature, it is generally applied in the financial domain to make predictions and classifications of stock price data. And there are many notions, which are making this method more promising for clustering domain, in the examined papers such as the idea of using fusions of different styles of charts and RGB charts.

On the other hand, as the experiments show, there are many hyperparameters which should be tuned when creating the visual representations of the data. For example, when creating a line chart, the width of the line, the sizes of the plots, the colouring style and many more decisions which should be made during the visualization process may play a crucial role to create a successful clustering model. Another downside of the idea is that the computational complexity and time will be much more than a standard application of clustering because of the two huge processes of data creation and feature extraction added before the clustering application.

In conclusion, it is a high potential idea to apply similar approaches to create better separated clusters of datasets or reach some solutions which cannot be reached by any other method. To improve performances in further studies, a better-defined loss function can be used, the datasets can be visualized using different plotting styles and the autoencoder model architecture can be designed better to create lower dimensional latent representations. Applying these improvements, the idea may be utilized better to have successful clustering results.

7. References

Bengio, Y., Courville, A., & Vincent, P. (2013). *Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828. doi:10.1109/tpami.2013.50

Best, R.J. (2004), "Market Based Management: Strategies for Growing Customer Value and Profitability", 3rd ed. Upper Saddle River, N.J.: Prentice Hall.

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). *Deep Clustering for Unsupervised Learning of Visual Features. Lecture Notes in Computer Science*, 139–156. doi:10.1007/978-3-030-01264-9_9

Chen, J.-H., & Tsai, Y.-C. (2020). *Encoding candlesticks as images for pattern classification using convolutional neural networks. Financial Innovation*, 6(1). doi:10.1186/s40854-020-00187-0

Hastie, T., Tibshirani, R., & Friedman, J. (2017). "*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*"

Hauser, W.J. (2007), "Marketing analytics: the evolution of marketing research in the twenty-first century", Direct Marketing: An International Journal, Vol. 1 No. 1, pp. 38-54.

Kim, T., & Kim, H. Y. (2019). *Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data. PLOS ONE*, 14(2), e0212320. doi:10.1371/journal.pone.0212320

Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., & Long, J. (2018). *A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture. IEEE Access*, 6, 39501–39514. doi:10.1109/access.2018.2855437

Ryu, S., Choi, H., Lee, H., & Kim, H. (2019). *Convolutional Autoencoder based Feature Extraction and Clustering for Customer Load Analysis. IEEE Transactions on Power Systems*, 1–1. doi:10.1109/tpwrs.2019.2936293

Unlu, E. (2020). *RGBSticks: A New Deep Learning Based Framework for Stock Market Analysis and Prediction. Journal of Soft Computing and Artificial Intelligence*, 1 (2), 20-28. Retrieved from https://dergipark.org.tr/tr/pub/jscai/issue/56697/798545