Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11594
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Using the Deep Constrained Clustering Approach to Create a Business Profile

Abdul Latif[1)], Sutarman[2)], Open Darnius[3)]
Universitas Sumatera Utara,Medan,Indonesia.

Abstract. Identification of customers in the business sector that really needs to be done as an evaluation of a business that is run so that it can continue to grow and be able to follow business developments in the same sector. The deep constraint clustering approach is used to cluster customers towards a business. In this study, a clustering of customers using rail mass transportation will be carried out. The results achieved are the formation of 6 clusters using trains be built. The result of research expected to be a consideration in improving services to the company.

Keywords:

INTRODUCTION

In the development of Information Technology at this time, data processing has been very massive carried out by carrying out several approaches and data construction (Rossmann & Van Beek, 1999). Data mining is the one of the processes of extracting the data from a large database to be used in data construction that involves many variables in it, the more data the more complex the value variations will be. The data mining approach is carried out by taking into account the constraints and factors that influence significantly because it will affect the segmentation of the data construction carried out. The formation of partitions or segmentation of a data construction really needs to be done and considered so as to be able to provide an overview of the variation of the data that is built. In its application the construction of customer segmentation in a business profile is very important to note because the quantity of consumers in a business is one of the benchmarks for the success of a business. The obstacle that needs to be studied is how to group customer data on the customer behavior of each product. Customer clustering is done to group customer data based on customer tendencies in choosing products, this is done to find out which products tend to be in demand and less attractive to customers. In terms of customer optimization, it is necessary to optimize the availability of products that are less attractive to customers because this will have an impact on the operational value that must be issued by a company, customer clustering will also have an impact on service improvement by implementing new strategies to serve and meet the needs of customers. In the clustering process, sometimes there will be outliers which are minority customers who are included in customer segmentation, the outlier data contained in the database needs to be analyzed so that an approach to calculating the distance between each cluster can be carried out. The Deep constrained clustering is the popular method of k-means and EM. The addition of the resulting constraint from the ground truth label allows a semi-supervised setting to increase accuracy when measured against a ground truth label (Zhang et al., 2021, 2019).

Deep constrained clustering combines equality constraints between several product pairs to find out which of several products in the data set are in the same group (positive constraint or must-link constraint) and which are not in the same group (negative constraint or constraint cannot). -link). Deep constrained clustering has been widely used in various real-world applications such as GPS-based map enhancement or landscape detection from hyperspectral data (Markos & James, 2020). On the other hand, in semi-supervised classification, learning involves using only a small amount of labeled data along with a large amount of data with unlabeled elements (Kipf & Welling, 2016). Basically, the pairwise constraints used in clustering cannot be directly converted into class labels, and it makes a conceptual difference between clustering and semi-supervised classification.

*Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v7i3.11594 p-ISSN : 2541-044X

The formulation of the problem in this research is how to use the Deep Constrained Clustering Approach to Create a Business Profile by clustering customers in a business so that it can be seen that customer clusters have the potential to improve services.

LITERATURE RIVIEW

Clustering

Cluster analysis is the work of grouping data (objects) based only on the information found in the data that describes the object and the relationship between them (Prasad & Balakrishnan, 2022). The goal is that the objects that join in a group are objects that are similar (or related) to each other and different (or unrelated) to objects in other groups. The greater the similarity (homogeneity) within the group and the greater the difference between the other groups, this concept will be discussed in the grouping. The purpose of working on data clustering can be divided into two, namely grouping for understanding and grouping for use. If the purpose is for understanding, the group in the form must capture the natural structure of the data, usually the grouping process in this goal is only a preliminary process to then proceed with core work such as summarization (mean, standard deviation), class labeling in each group for later used as training data classification, and so on. Meanwhile, if for use, the main purpose of grouping is usually to find a prototype group that is most representative of the data, providing an abstraction of each data object in the group where a data is located. There are many methods of grouping (clustering) that have been developed by experts. Each method has its own character, advantages, and disadvantages. Clustering methods are grouped into four categories: partitioning methods, hierarchical methods, density-based methods, and grid-based methods. method) (Kumar et al., 2003).

Partitioning Method

As the name implies, this method works by dividing or partitioning data into a number of groups. This method is also known as the center-based method or representative-based method (Rozita et al., 2014) because it works by determining cluster centers, where the cluster center can be an average, mode, or a representative object of all objects in a cluster based on a certain size. As a formulation. Suppose you have a data set D containing n objects in Euclidean space. This method divides *n* objects into k clusters, $C_1, C_2, \ldots, C_k$, without any overlapping objects, meaning $C_i \subset D$ and $C_i \cap C_j = for 1 \leq i, j \leq k$. In this method the formulation of Sum of Square Error (SSE) written as:

$$\text{SSE} = \sum_{i=1}^{k} \sum_{p \in C_i} dist\ (p, c_i)$$

In this method, a cluster C_i is represented as a centroid or conceptually as a cluster center point. Censtroid can be the average (mean) or mode (modes) of all objects or data points in a cluster or it can be medoid (an object or representative point that represents all objects in a cluster). The difference between an object p in cluster C_i and the centroid c_i is calculated using the Euclidean distance *dist(p, c_i)*. C_i cluster quality can be measured using variations within the cluster, namely the sum of squared error (SSE) between all objects in cluster C_i and the centroid c_i.

Deep Constrained Clustering

Constrained clustering aims to find clusters that meet user-specific constraints (Padilla & Blanco, 2020). User constraints can be classified into: cluster-level constraints, specifying requirements on cluster or instance-level constraints, and specifying requirements on instance pairs. An instance-level constraint (also called a pairwise constraint) is a constraint on pairs of instances (Ferracuti et al., 2019). There are two types of instance-level constraints that were first introduced by, bound and non-linkable constraints. The mandatory link constraint between two objects o_i, o_j , denoted by ML(o_i, o_j), states that both objects o_i and o_j must be in the same cluster. On the other hand, the constraint cannot link between o_i and o_j , denoted by CL(o_i, o_j), meaning that the two objects cannot be in the same cluster. (Santos et al., 2018), instance-level constraints have the following:

*Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v7i3.11594 p-ISSN : 2541-044X

1. The Transitive-constraint: Given CC_a and CC_b as the connected components, let oi and the value of oj be objects in CC_a and CC_b, respectively. CCb, then (Sánchez et al., 2015):
$$ML\ (o_i, o_j), o_i\ \in\ CC_a, o_j\ \in\ CC_b\ \Rightarrow\ ML\ (o_x, o_y), \forall o_x, o_y : o_x\ \in\ CC_a, o_y\ \in\ CC_b$$
2. A cannot-link problem may occur. Suppose CC_a and CC_b as a connected components, let oi and oj be objects in CC_a and CC_b respectively, then:
$$CL\ (o_i, o_j), o_i\ \in\ CC_a, o_j\ \in\ CC_b\ \Rightarrow\ CL\ (o_x, o_y), \forall o_x, o_y : o_x\ \in\ CC_a, o_y\ \in\ CC_b$$

Constraint-Based Method

The part modified called COP-Kmeans to integrate bound and non-linkable constraints. The main modification is that on each iteration, the object is updated so that no constraint is contravened. The drawback of this approach is that it tries to satisfy all the constraints but the algorithm does not provide any backtracking technique. As a result, the algorithm may fail to find the existing partition. It happens when there are many obstacles, especially the cannot-link constraint:

$$m_c = \frac{\sum_{o_i \in C_c} o_i}{|C_c|}$$

thus, the objects move to closets centroid respectively to minimize the vector quantization error (VQE), written as:

$$\text{VQE} = \frac{1}{2}\sum_{c\in[1,k]}\sum_{o_i \in C_c}\|m_c - o_i\|^2$$

The algorithm modifies the error function to punish the contravened constraint, defined vector quantization error finite $CVQE_c$ of a cluster c_c by:

$$CVQE_c = \frac{1}{2}\sum_{o_i \in C_c}(m_c - o_i)^2 + \frac{1}{2}\sum_{o_i \in C_c, ML\ (o_i,o_j) o_j \in C'_c, c' \neq c}\|m_c - m'_c\|^2 + \frac{1}{2}\sum_{o_i \in C_c, CL\ (o_i,o_j) o_j \in C_c}\|m_c - m_{h(c)}\|^2$$

METHODS

Customer Segmentation Cluster Analysis

Customer segmentation is the process by which a company divides its base customers into different groups based on some shared characteristics (Jardim & Mora, 2022). This is done so that different groups can be analyzed and marketing can be tailored to these groups based on their preferences to increase sales and customer relationships.

Cluster analysis is a branch of the mathematical discipline that aims to grouping similar data into subsets called cluster. Part of the data must have the same characteristics so that the cluster data have similarities from other clusters. The essence of cluster analysis is to determine the distance from the data point which is usually described as N ×N (where N is the number of data) the dissimilarity of the matrix D which has a member consists of the distance between to observation. It is assumed that a matrix:

$$\boldsymbol{X} = \{x_{ij}\}_{1\leq i\leq n, 1\leq j\leq p}$$

Where is the number of data and p is the dimension of each data which has a dissimilarity $d_j(x_{ij}, x_{i'j})$ between the values of j . The value of dissimilarity of each data is defined as:

$$D(x_i, x_{i'}) = \sum_{j=1}^{p} w_j d_j(x_{ij}, x_{i'}), \sum_{j=1}^{p} w_j = 1$$

RESULT AND DISCUSSION

Problem Description

A railway company working in the transportation sector transports passengers from station A to station B. Passengers who are escorted are very heterogeneous, in the survey that carried out there were a total of 8.068 passengers (data attached). Each passenger is distinguished by ID, gender, marital status, age, status study, work, work experience, expenses, number of families, and segmentation train carriage. To improve management services, it is necessary to make observations passengers and form customer clustering so that it can be known which customers have the same

*Corresponding author

tendency to use train mass transportation facilities fire. The formation of customer clusters needs to pay attention to the constraints that have been obtained from the data customers, the complexity of the constraints will give the average customer value using rail mass transportation in this case the management will do customer clustering into 6 clusters (k = 6). By knowing the customer cluster It is hoped that the management will be able to provide priority service to passengers in the future.

Optimal Strategy Algortihm

The optimal strategy of business profile have two strategy of training and effectively such that create mini batches for training. The example of the difficulty or global size constraint, treat their costs function as addictive costs so there is no additional branches need to be created. Costs branches use more complex costs functions because the constraints defines on pairs and even triples instances. So create a branch another costs that contains a pairwise costs or a triplet costs $ł_p$ to help the network tune an embed that satisfies this stronger constraint. For each type constraint, a mini-batch is created consisting of only instances that have that type of constraint. For each instance of a constraint type, given an instance that is restricted over the network, computes costs, calculates the change in weight but does not adjust the weight. Algorithm will summing the weight adjustments for all instances of the constraint in the mini-batch and then adjust weights. Therefore the branch costs method is an example of weight updating batch as standard in DL for stability reasons. The whole training procedure is summarized in the following Algorithm:

Algorthm. The Framework of Deep Constrained Clustering
Input. sum of X: data, m: epoch maximum, k: sum of cluster, N: sum of stage, Nc: sum of constraint of the stage
Output. The cluster of the population thus the value of cluster can be evaluate in developing the bussiness
IF dataset is numeric **choose numeric**
ELSE IF dataset is character **choose character**
Transform character into numerical/dummy
calculate lC and lR
calculate lI and lG
calculate sum of the costumer as lc+lR+(lI||lT)
evaluate the costumer network based on the sum of costumer
ELSE
FOR segment = 1 to *NAs* do
calculate lPandlT
calculate total of costumer segmentation
END FOR
Calculate the optimal value of cluster
END

Algorithm Simulation Using R

In the simulation used, the author utilizes the R Programming work environment to perform data normalization, calculations, and row transformations into appropriate variables for each constraint in the pre-processing stage. Furthermore, the author uses Microsoft Excel as a simulation output in .csv format which will then select the best customer cluster and eliminate variables that are at the furthest distance from the cluster.

*Corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11594
e-ISSN : 2541-2019
p-ISSN : 2541-044X

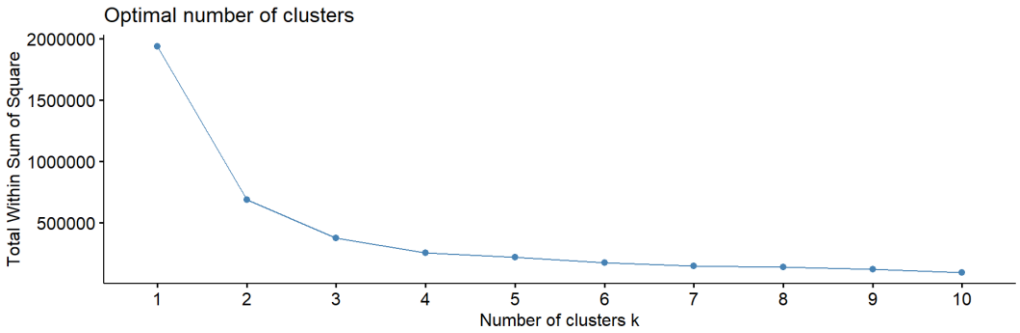

FIGURE 1. Optimal Cluster

Based on the computational results of the R program, the optimal number of clusters is 10 clusters, because management only wants 6 clusters it needs to be done minimization of clusters 6 out of 10 by using the syntax clusters <- kmeans (dataset, centers = 6, iter.max = 10. By exporting data into the form .csv write.csv (clusters$centers, file = "clustering.csv") will be shown:

Cluster	ID	Age	Work_Experience	Family_Size	Gender_Female	Ever_Married_No	Graduated_No	Profession_Doctor	Profession_Engineer	Profession_Entertainment	Profession_Executive	Profession_Healthcare	Profession_Homemaker	Profession_Lawyer	Profession_Marketing	Spending_Score_Average	Spending_Score_High
1	0.071	0.120	(0.054)	(0.146)	(0.051)	(0.831)	(0.184)	3.211	(0.309)	(0.372)	(0.286)	(0.440)	(0.165)	(0.284)	(0.190)	0.882	(0.180)
2	0.010	(0.042)	0.043	(0.052)	(0.372)	(0.044)	0.005	(0.311)	(0.309)	2.691	(0.286)	(0.440)	(0.165)	(0.284)	(0.190)	0.203	(0.308)
3	(0.003)	(1.023)	(0.013)	0.608	(0.066)	0.984	0.592	(0.311)	(0.309)	(0.372)	(0.283)	2.234	(0.165)	(0.281)	(0.190)	(0.505)	(0.344)
4	0.013	0.770	(0.259)	(0.067)	(0.213)	(0.791)	(0.273)	(0.311)	(0.297)	(0.372)	0.488	(0.427)	(0.155)	0.499	(0.160)	0.306	0.570
5	(0.013)	(0.266)	0.340	(0.261)	0.505	0.394	(0.032)	(0.311)	0.829	(0.372)	(0.259)	(0.440)	0.437	(0.276)	0.478	(0.212)	`(0.312)
6	(0.088)	(0.758)	0.021	0.131	(0.013)	1.096	0.330	3.211	(0.309)	(0.372)	(0.286)	(0.440)	(0.165)	(0.284)	(0.190)	(0.577)	(0.413)

FIGURE 2. Costumer Segmentation

Based on the costumer segmentation simulation, we have to eliminate the same value that have no uniqe value of each cluster, so the eliminate segmentation shown at the figure below Based on simulation results with deep constraint clustering algorithm for business profile.

Cluster	Age	Work_Experience	Family_Size	Gender_Female	Ever_Married_No	Graduated_No	Profession_Doctor	Profession_Engineer	Profession_Entertainment	Profession_Executive	Profession_Healthcare	Profession_Homemaker	Profession_Lawyer	Profession_Marketing	Spending_Score_Average	Spending_Score_High
Single Doctor/ Healthcare/Entertainment	0.120	(0.054)	(0.146)	(0.051)	(0.831)	(0.184)	3.211	(0.309)	(0.372)	(0.286)	(0.440)	(0.165)	(0.284)	(0.190)	0.882	(0.180)
Single Male Entertainment /Healthcare	(0.042)	0.043	(0.052)	(0.372)	(0.044)	0.005	(0.311)	(0.309)	2.691	(0.286)	(0.440)	(0.165)	(0.284)	(0.190)	0.203	(0.308)

*Corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Old Single Healthcare/Entertainment	(1.023)	(0.013)	0.608	(0.066)	0.984	0.592	(0.311)	(0.309)	(0.372)	(0.283)	2.234	(0.165)	(0.281)	(0.190)	(0.505)	(0.344)
Old Single Lawyer/ Executive /Healthcare/ Entertainment	0.770	(0.259)	(0.067)	(0.213)	(0.791)	(0.273)	(0.311)	(0.297)	(0.372)	0.488	(0.427)	(0.155)	0.499	(0.160)	0.306	0.570
Female Married Engineer /Marketing/ Healthcare/Home Maker /Entertainment	(0.266)	0.340	(0.261)	0.505	0.394	(0.032)	(0.311)	0.829	(0.372)	(0.259)	(0.440)	0.437	(0.276)	0.478	(0.212)	(0.312)
Old Single Doctor /Healthcare/ Entertainment	(0.758)	0.021	0.131	(0.013)	1.096	0.330	3.211	(0.309)	(0.372)	(0.286)	(0.440)	(0.165)	(0.284)	(0.190)	(0.577)	(0.413)

The results showed that the 6 clusters produced showed heterogeneity of customers who using rail transportation. The 6 selected clusters are:

1. Cluster 1: Young Single Doctor / Healthcare/Entertainment
2. Cluster 2: Single Male Entertainment /Healthcare
3. Cluster 3: Old Single Healthcare / Entertainment
4. Cluster 4: Old Single Lawyer / Executive / Healthcare /Entertainment
5. Cluster 5: Female Married Engineer / Marketing /Home Maker /Healthcare / Entertainment
6. Cluster 6: Old Single Doctor / Healthcare/ Entertainment

With the formation of the passenger cluster into 6 clusters, the management can consider the service of train carriages that suit the needs of the customers, passenger segmentation is done by taking the average value of passengers in the same cluster.

CONCLUSION

Based on the research that has been done, it can be concluded that the method approach deep constraint clustering for business profiling can be done by the formation of segmentation of passenger data based on the specified constraints. Results The results obtained are in the form of passenger clusters that have a tendency to use public transportation modes rail mass transportation. By knowing the characteristics of passengers in each cluster, it can be be a consideration for the services to be provided so as to be able to provide improvement of business services that have been carried out.

REFERENCES

Ferracuti, N., Norscini, C., Frontoni, E., Gabellini, P., Paolanti, M. & Placidi, V. (2019). A business application of RTLS technology in Intelligent Retail Environment: Defining the shopper's preferred path and its segmentation. *Journal of Retailing and Consumer Services*, *47*, 184–194.

Jardim, S. & Mora, C. (2022). Customer reviews sentiment-based analysis and clustering for market-oriented tourism services and products development or positioning. *Procedia Computer Science*, *196*, 199–206.

Kipf, T. N. & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *ArXiv Preprint ArXiv:1609.02907*.

Kumar, S., Loui, A. C. & Hebert, M. (2003). An observation-constrained generative approach for probabilistic classification of image regions. *Image and Vision Computing*, *21*(1), 87–97.

Markos, C. & James, J. Q. (2020). Unsupervised deep learning for GPS-based transportation mode

*Corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

identification. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–6.

Padilla, A. O. & Blanco, J. C. (2020). Shopping centre clusters: Competition or synergies? The case of the region of Murcia (Spain). *Journal of Retailing and Consumer Services*, *52*, 101867.

Prasad, M. V. N. K. & Balakrishnan, R. (2022). Spatio-Temporal association rule based deep annotation-free clustering (STAR-DAC) for unsupervised person re-identification. *Pattern Recognition*, *122*, 108287.

Rossmann, M. G. & Van Beek, C. G. (1999). Data processing. *Acta Crystallographica Section D: Biological Crystallography*, *55*(10), 1631–1640.

Rozita, A. L., Zana, A. A. N., Khairulzaman, H. & Norlizah, A. H. (2014). Impact of sport complex services towards costumer behaviour in Terengganu. *Procedia-Social and Behavioral Sciences*, *153*, 410–418.

Sánchez, E. M., Clempner, J. B. & Poznyak, A. S. (2015). Solving the mean–variance customer portfolio in Markov chains using iterated quadratic/Lagrange programming: A credit-card customer limits approach. *Expert Systems with Applications*, *42*(12), 5315–5327.

Santos, P. M., Kholkine, L., Cardote, A. & Aguiar, A. (2018). Context classifier for position-based user association control in vehicular hotspots. *Computer Communications*, *121*, 71–82.

Zhang, H., Basu, S. & Davidson, I. (2019). A framework for deep constrained clustering-algorithms and advances. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 57–72.

Zhang, H., Zhan, T., Basu, S. & Davidson, I. (2021). A framework for deep constrained clustering. *Data Mining and Knowledge Discovery*, *35*(2), 593–620.

*Corresponding author