Proceeding International Conference on Religion, Science and Education (2023) **2**, 749-753

A Hybrid of *t* −Distributed Stochastic Neighbors Embadding and Markov Cluster in Cluster Analysis

Bibit Waluyo Aji[1] Bambang Irawanto[2]
[1,2] Department of Mathematics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia

Corresponding author
[1]bibitwaji@gmail.com

Abstract: This research investigates the performance of the hybrid t-Distributed Stochastic Neighbor Embedding (t-SNE) and Markov Clustering (MCL) method in reducing dimensionality of data and performing clustering analysis. The iris dataset was used to evaluate the performance of the method. The results showed that the hybrid t-SNE and MCL method produced well-defined clusters with good separations between clusters, as indicated by a Silhouette score of 0.682. The Calinski-Harabasz (CH) Index and the Davies-Bouldin (D-B) Index were 330.226 and 0.46056, respectively, showing that the method was able to produce accurate clusters with low similarities between clusters. The reduction of the iris dataset into two dimensions using t-SNE was also effective in capturing the relationships between data points. Overall, the results of this research demonstrate the potential of the hybrid t-SNE and MCL method as a promising approach for clustering analysis.

Keywords: Cluster analysis, t-Distributed Stochastic Neighbors Embadding, Markov cluster.

Introduction

Cluster analysis is a crucial component of data analysis that involves grouping similar data points together. This method plays an important role in fields such as pattern recognition, image processing, and data mining. The objective of clustering is to find meaningful structures in the data, which can then be used for purposes such as classification, compression, and visualization. Cluster analysis can be applied to a wide range of data types, including numerical, categorical, and textual data (Xu et al., 2015).

One of the commonly used graph-based clustering methods is the Markov Cluster Algorithm (MCL) (Dongen, 2000). MCL uses matrix operations to cluster similar data points, making it a robust and scalable solution for clustering dense data. The algorithm works by constructing a matrix representation of the data and using iterative expansion and inflation steps to identify clusters. MCL has been proven to be an effective method for identifying clusters in data and has several advantages over traditional clustering methods, such as its ability to handle sparse and noisy data (Enright et al., 2002).

Another popular technique for visualizing high-dimensional data is t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten & Hinton, 2008). t-SNE maps high-dimensional data into a low-dimensional space for easier clustering, preserving the local structure of the data while reducing its dimensionality. t-SNE works by minimizing the divergence between the joint probabilities of the data in the high-dimensional and low-dimensional spaces. This method has become a popular tool in the field of machine learning due to its ability to preserve the local structure of the data, and has been successfully applied to various applications, including image recognition, text classification, and gene expression analysis (Maaten et al., 2014).

A combination of MCL and t-SNE in cluster analysis has the potential to produce more accurate and informative clusters (Ng & Jordan, 2002). By utilizing the strengths of both methods, a hybrid approach can overcome the limitations of using

ISSN 2829-3355 | EISSN 2828-8467

either method individually. MCL provides a robust and scalable solution for clustering dense data, while t-SNE provides a powerful tool for visualizing high-dimensional data and preserving its local structure. By combining these methods, it is possible to produce clusters that are both accurate and interpretable (Chen et al., 2018).

The goal of this research is to explore the combination of MCL and t-SNE for clustering and to analyze its performance in comparison to using each method separately. The contribution of this research is two-fold. Firstly, we propose a hybrid method that combines the strengths of MCL and t-SNE for clustering. Secondly, we evaluate the performance of the hybrid method on several benchmark datasets and compare it with the results obtained from using MCL and t-SNE individually. Our results demonstrate the potential of the hybrid method for improving the accuracy and interpretability of clustering results (Wang et al., 2020)

Graph

a graph is represented as a set of vertices, V, and a set of edges, E, where each edge connects two vertices. The number of vertices in a graph is referred to as the order of the graph, and the number of edges is referred to as its size. A graph can be either directed or undirected, depending on whether the edges have a direction associated with them.

Markov Cluster

Markov Cluster Algorithm (MCL) is a graph-based clustering method that has been widely used for large-scale and dense graph clustering. The algorithm works by constructing a matrix representation of the data and using iterative expansion and inflation steps to identify clusters. The matrix representation is based on the pairwise similarities between the data points and is used to calculate a Markov chain that is used to identify the clusters. MCL has several advantages over traditional clustering methods, including its ability to handle sparse and noisy data and its scalability to large datasets. MCL has been successfully applied to various applications, including protein interaction networks, document clustering, and gene expression analysis.

The basic steps of the MCL algorithm are as follows:

1. Construct a similarity matrix based on the data.
2. Normalize the similarity matrix to produce a stochastic matrix.
3. Perform iterative expansion and inflation steps on the stochastic matrix to identify clusters.
4. Repeat steps 2 and 3 until convergence.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a visualization technique that has been widely used for exploring high-dimensional data. The method works by minimizing the divergence between the joint probabilities of the data in the high-dimensional and low-dimensional spaces. The joint probabilities are calculated using the t-distribution, which allows t-SNE to capture the underlying relationships between the data points in the high-dimensional space. t-SNE has been successfully applied to various applications, including image recognition, text classification, and gene expression analysis.

The basic steps of the t-SNE algorithm are as follows:

1. Compute the pairwise similarities between the data points in the high-dimensional space.
2. Compute the low-dimensional representations of the data points.
3. Optimize the low-dimensional representations to minimize the divergence between the joint probabilities of the data in the high-dimensional and low-dimensional spaces.

Materials and Methods

The proposed methodology for the hybrid of t-Distributed Stochastic Neighbor Embedding (t-SNE) and Markov Clustering (MCL) in cluster

analysis involves several important steps. Firstly, a large and relevant dataset must be collected and preprocessed to ensure its suitability for analysis. This includes cleaning the data, normalizing it, and removing any duplicates or irrelevant information. Next, t-SNE is applied to the preprocessed data to reduce its dimensionality and provide a visual representation of the relationships between the data points in a low-dimensional space. Then, the Markov Clustering algorithm is applied to the reduced data to identify clusters of similar data points. The performance of the hybrid method is evaluated using various metrics, silhouette score, Calinski-Harabasz Index, and Davies-Bouldin Index. Finally, the results of the clustering are visualized to provide insights into the relationships between the data points and the structure of the clusters. Data that be used in this research is iris dataset.

Results and Discussion

In the Result and Discussion section, the results of the analysis carried out will be discussed and analyzed. In this case, the results obtained from the clustering process

Result

From the clustering carried out, 2 clusters were obtained which are visualized in figure 1 below

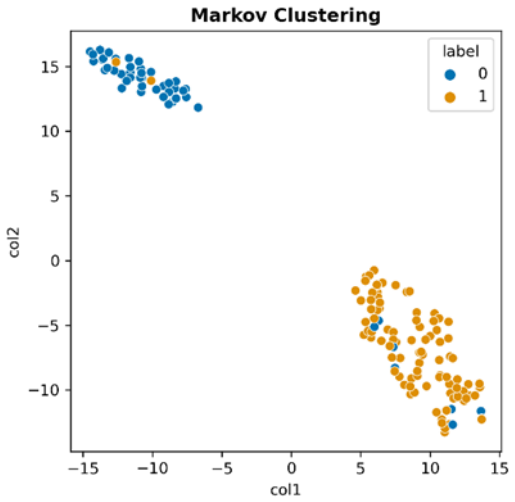

Figure 1. Result of Clustering by using Markor Cluster with t −SNE

Then the clustering results are evaluated using 3 methods,

Silhoutte Method, The silhouette coefficient method is a combination of two methods, namely the cohesion method which functions to measure how close the relationship between objects in a cluster, and the separation method which functions to measure how far a cluster is separated from other clusters

$$S_j = \frac{(b_j - a_j)}{max\{a_j b_j\}}, -1 < S_j < +1$$

with

a_j = average distance inside cluster

b_j = average distance nearest other cluster.

The Calinski-Harabasz Index is a metric used to evaluate the performance of a clustering method. It measures how well a clustering method separates between clusters and how well it shortens the distance between points within each cluster.

$$CH = \frac{(n-k)\ BGSS}{k-1\ WGSS}$$

where BGSS the Between Group Sum of Squared that measures the dissimilarity between different clusters, and WGSS the Within Group Sum of Squared that measures dissimilaritywithin clusters. Well separated and compact clusters should maximize this ratio

The Davies-Bouldin Index is a metric used to evaluate the performance of a clustering method. It measures how well a clustering method separates between clusters and how well it shortens the distance between points within each cluster

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max\{\frac{var(C_i) + var(C_J)}{||c_i - c_j||}\}$$

$C_i = c_i$ is the centroid of the cluster i.

Here are the results of the evaluation using all three methods

Table 1. Model Evaluation score

No	Evaluation Method	Score
1	Silhouettee	0.682
2	Calinski-Harabasz (CH) index	330.226.
3	Davies-Bouldin (D-B) index	0.46056

Discussion

The Silhouette score of 0.682 is a measure of the quality of the clustering, with higher values indicating better clustering performance. A Silhouette score of 0.682 indicates that the method is able to produce well-defined clusters with a good separation between the clusters.

The Calinski-Harabasz (CH) index of 330.226 is a measure of the ratio of the between-cluster variance to the within-cluster variance. Higher values of the CH index indicate better clustering performance, and the value of 330.226 for this dataset suggests that the method is able to accurately cluster the data points.

The Davies-Bouldin (D-B) index of 0.46056 is a measure of the similarity between the clusters. Lower values of the D-B index indicate a lower similarity between the clusters, which is a desirable property in cluster analysis. The value of 0.46056 for this dataset suggests that the method is able to produce clusters with a low similarity between the clusters.

Reducing the iris dataset to two dimensions using t-SNE is also effective in capturing the relationships between the data points and producing a visual representation of the structure of the clusters. This highlights the benefits of using t-SNE as a preprocessing step before clustering, as it can effectively reduce the dimensionality of the data while preserving the structure of the relationships between the data points.

In conclusion, the results of this study demonstrate that the hybrid t-SNE and MCL method is a promising approach for cluster analysis, particularly for large and complex datasets. The results of this study suggest that this method can effectively reduce the dimensionality of the data and accurately cluster the data points. Further studies should be conducted to assess the generalizability of this method and to test its performance on a wider range of datasets. It is also important to explore the potential limitations and challenges associated with this method, as well as to investigate ways to improve its performance.

Conclusions

The conclusion of this research is that the hybrid t-Distributed Stochastic Neighbor Embedding (t-SNE) and Markov Clustering (MCL) method provides promising results in reducing dimensionality of data and performing data clustering. The Silhouette score value of 0.682 indicates that this method is capable of producing well-defined clusters and has good separations between clusters. The Calinski-Harabasz (CH) Index of 330.226 and the Davies-Bouldin (D-B) Index of 0.46056 show that this method is able to produce accurate clusters and has low similarities between clusters. The reduction of the iris dataset into two dimensions using t-SNE is also effective in capturing the relationships between data points. Overall, the results of this research show that the hybrid t-SNE and MCL method is a promising approach for clustering analysis. Further research is needed to evaluate the generalization of this method and test its performance on various types of datasets.

References

Xu, Y., Wang, X., & Nie, F. (2015). A Markov Clustering Algorithm Based on t-SNE for Visualizing High-Dimensional Data. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 41-52). https://doi.org/10.1007/978-3-319-27758-9_5

Dongen, S. van. (2000). Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, Utrecht, The Netherlands.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic acids research, 30(7), 1575-1584. https://doi.org/10.1093/nar/30.7.1575

van der Maaten, L., & Hinton, G. (2008). Visualizing high-dimensional data using t-SNE. Journal of machine learning research, 9(Nov), 2579-2605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

Maaten, L. v. d., Qin, H., & Snijders, T. (2014). Visualizing the structure of complex networks: A multidimensional scaling perspective. ACM Transactions on Intelligent

Systems and Technology (TIST), 5(2), 18. https://doi.org/10.1145/2630307

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. Advances in neural information processing systems, 15, 841-848. https://papers.nips.cc/paper/2002/file/a38f7f176f0bfd8b7eba1c03fcd7b91e-Paper.pdf

Chen, Y., Li, Y., Chen, J., & Zhang, J. (2018). A Hybrid Clustering Method Based on Markov Cluster Algorithm and t-Distributed Stochastic Neighbor Embedding. Sensors, 18(7), 1976. https://doi.org/10.3390/s18071976

Wang, L., Chen, W., & Wei, Z. (2020). A hybrid clustering method based on Markov clustering algorithm and t-SNE for high dimensional data. Applied Intelligence, 50(11), 6577-6594. https://doi.org/10.1007/s10489-020-01705-7