Data Clustering Based on the Modified Inversion Formula Density Estimation

INTRODUCTION

Data research is widely used in various fields such as business, production, online trade, consumer services, and other fields. Due to such a large data mining application, the field is receiving much attention. Data clustering is an unsupervised type of machine learning that is also widely used in data mining. In data clustering, the main goal is to divide objects into separate, unknown groups to have as many similar objects as possible in one group. Making such groups allows you to find hidden relationships between data. Data clustering is applied in bioinformatics, feature selection, and pattern recognition. Although there are many different methods in data clustering, it is a complex task. Due to different data structures, different clustering methods work well only under certain conditions, so the need for these methods remains high. One of the most used data clustering methods is the k-means method, which is relatively simple but can work effectively in good conditions. Most clustering methods perform poorly in the presence of outliers in the data, and the previously mentioned k-means method suffers from this drawback, as do GMM, BGMM, and other methods. Recently, various researchers have been paying much attention to different density estimation methods, as well as robust modifications of these methods, such as soft-constrained neural networks and others. Due to such a demand for density estimation, this paper aims to evaluate the accuracy of a new clustering method based on modified inversion formula density estimation. The results show that this developed method is competitive compared to the current most popular methods (K-means, GMM, BGMM).

CLUSTERING BASED ON THE MODIFIED INVERSION FORMULA

In the formula for calculating the density estimate, construct the estimate of the characteristic function as a union of the characteristic functions of a mixture of Gaussiandistributions and a uniform distribution with corresponding a priori probabilities.

$$\hat{\psi}_\tau(u) = \sum_{k=1}^{\hat{q}_\tau} \hat{p}_{k,\tau} e^{iu\hat{m}_{k,\tau} - u^2\hat{\sigma}^2_{k,\tau}/2} + \hat{p}_{0,\tau}\frac{2}{(b-a)u}\sin\frac{(b-a)u}{2}\cdot e^{\frac{iu(a+b)}{2}}$$

Input: Data set X= [X1,X2, . . . , Xn], cluster number K
Output: C1, C2, _, Ct and M, p, R
Possible initiation of mean vector:
(1) random uniform initialization
(2) k-means
(3) random point initialization
Generate a T matrix. The set T is calculated when the design directions are evenly spaced onthe sphere.
For i = 1: t do
Density estimation for each point and cluster based on formula
Update M, pk, R matrices
End
Return C1, C2, _, Ct and M, pk, R

RESULTS, LIMITATIONS, FUTURE RESEARCH

Based on the clustering results, it can be observed that the CBv2 method works the best with generated data with noise in all datasets (0.5%, 1%, 2%, and 4% noise). Based on the accuracy metric with all of these datasets, accuracy was higher than 0.995. The interesting point is that a new method based on the inversion formula can cluster the data even if data do not have outliers; one of the most popular, for example, is the Iris data set. When we compared the accuracy results in other datasets, it can be mentioned that the CBMIDE method achieved 0.955 accuracies on the Iris dataset compared with the second-best GMM method with 0.953 accuracy; using the ARI metric for this dataset, CBMIDE methods as well showed better results compared with other methods. Additionally, it is notable that the CBMIDE method has a lower standard deviation than other methods used in this research. It is worth mentioning that this method also has limitations. Based on the experimental study, this one method in the current state can not work with higher dimensional data (d > 15). This occurs due to T matrix generation; as dimensions grow, finding a suitable T matrix becomes harder. This one will be solved in future versions of the model; we will present more about it in future work. The future direction of the newly created method is this method application for deep clustering. It can be seen that CBMIDEv1 and CBMIDEv2 methods do not work well with higher-dimension data. Due to that, the deep clustering method with an encoder structure could solve this problem.

AUTHORS

Mantas Lukauskas*, Tomas Ruzgas
***mantas.lukauskas@ktu.lt**
Microsoft Teams link: Meeting ID: 383 559 040 865 Passcode: zkoSFb