Analysis of Clustering Methods Performance Across Multiple Datasets

kaunas university of technology
ktu 1922

Abstract

As the amount of data increases each year, these amounts of data become increasingly difficult to analyze. Currently, a variety of different machine learning algorithms are proposed for data analysis to help make different versions, and research and other activities require solutions. One of the most commonly used forms of unsupervised learning is clustering. Clustering is often described as a particular process that seeks to find data contained in hidden relationships. It is unnecessary to know the class in advance to find these connections, which allows the data in the main groups to be distinguished. Data clustering can be performed using various methods, but they are all divided into four main groups: partitioning methods, hierarchical methods, density-based methods, and grid-based methods. This work aims to compare different groups of clustering methods and particular methods using different data sets and evaluate their performance. This work also seeks to include methods that are better known to everyone and much less commonly used. Finally, this work will help to provide some guidance on when specific methods are best suited.

Data and methods

Three different types of data were used in this study: Real, Artificial, and Generated. Real and Artificial datasets have been collected from a variety of sources and are all commonly used datasets in clustering studies. Generated data sets are composed by the authors themselves, which are generated with various parameters such as density, overlap, and so on. In total, the study is conducted with 113 data sets.

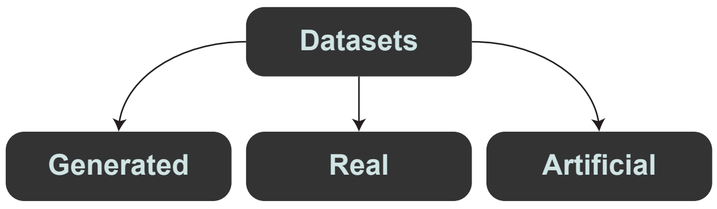

Figure 1. Data types user in research

Different types of data clustering algorithms were used in the study. The aim of this study was to make the widest possible comparison of different clustering algorithms, therefore K-means, DBSCAN and other methods were included. several models of mixtures have also been evaluated, namely Gaussian, Bayesian Gaussian and Von Mises Fisher Mixture. Total number of **65 201** models were created in one cycle (datasets x methods x params)

Results

The table below shows only some of the results of the studies. It can be seen that there is no existing algorithm that would work best in all cases. All algorithms were evaluated by 18 different metrics and 3 main metrics are presented NMI - Normalized Mutual Information, ARI - Adjusted Rand Index and AMI - Adjusted Mutual Information.

Table 1. Sample of research results (Only 6 datasets with 12 methods)

Models	FLAME			GLASS			ECOLI			WINE			IRIS			CANCER		
	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI	ARI
A-DBSCAN	**1,000**	**1,000**	**1,000**	0,378	0,392	0,252	0,4988	0,4895	0,499	0.5897	0.5897	0,039	0,7336	0,7315	0,568	0,265	0,261	0,220
DBSCAN	**1,000**	**1,000**	**1,000**	0,391	0,417	0,244	0,4988	0,4895	0,499	0,6162	0,6120	0,577	0,7336	0,7315	0,568	0,265	0,261	0,220
OPTICS	0,9918	0,9917	0,995	0,284	0,307	0,104	0,5069	0,4979	0,461	0,7565	0,7539	0,726	0,7336	0,7315	0,568	0,141	0,138	0,117
BIRCH	0,025	0,019	0,022	**0,457**	**0,484**	**0,299**	**0,7339**	**0,7225**	**0,778**	0,9385	0,9378	0,950	0,7837	0,7809	0,644	0,566	0,565	0,689
FINCH	0,0920	0,0860	0,024	0,292	0,327	0,165	0,6497	0,6338	0,649	0,7527	0,7501	0,741	0,7776	0,7748	0,744	0,585	0,584	0,706
CURE	0,2344	0,2289	0,188	0,385	0,414	0,247	——	——	——	——	——	——	0,8850	0,8836	**0,903**	0,349	0,348	0,423
K-MEANS	——	——	——	0,294	0,324	0,166	0,6281	0,6120	0,439	0,8529	0,8513	0,868	0,7419	0,7386	0,716	0,623	0,622	0,730
GMM	——	——	——	0,325	0,353	0,189	0,6425	0,6263	0,641	0,8770	0,8757	0,896	**0,8996**	**0,8984**	**0,903**	**0,668**	**0,667**	**0,780**
BGMM	——	——	——	0,324	0,352	0,192	0,6494	0,6388	0,681	**0,9537**	**0,9532**	**0,963**	0,7649	0,7619	0,684	0,653	0,652	0,767
ELASTIC	0,1402	0,1332	0,139	0,236	0,267	0,161	0,4061	0,3805	0,262	0,5893	0,5849	0,505	0,7032	0,6988	0,560	0,262	0,261	0,356
AP	0,4347	0,4160	0,151	0,303	0,385	0,129	0,5519	0,5122	0,226	0,5570	0,5385	0,323	0,6433	0,6317	0,581	0,271	0,256	0,058
AGG	**1,000**	**1,000**	**1,000**	0,413	0,441	0,262	0,7255	0,7139	0,748	0,9086	0,9076	0,931	0,7837	0,7809	0,719	0,416	0,416	0,538
SPECTRAL	——	——	——	0,301	0,332	0,176	0,6171	0,6005	0,585	0,9087	0,9077	0,931	0,7771	0,7743	0,338	0,554	0,553	0,623

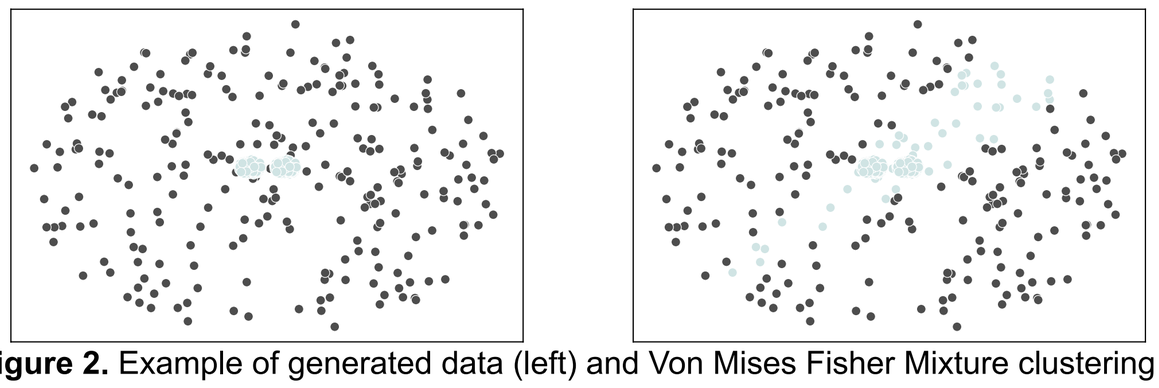

Figure 2. Example of generated data (left) and Von Mises Fisher Mixture clustering (right)

Conclusion

Based on the presented and additional results, it can be observed that there is no single method that would work successfully in all cases. DBSCAN and Adaptive DBSCAN perform well in complex shape clusters. Successful clustering of the BIRCH method is also observed for multidimensional data. It is also worth noting that the blend models have relatively good results in most clustering cases.

Future Research

This study is the second step in an overall larger study of clustering algorithms. There are currently contacts with foreign researchers regarding the inclusion of their clustering methods in the comparative study, so the study will be expanded with new methods. The results also showed that the mixture models performed well, so a clustering method based on MIDE (modified inversion formula density estimation) is currently being developed and will be included in these studies. Also taking into account that the generated data with different parameters allow successful isolation of the methods will be carried out further simulation study.

1. Simulation study with newly generated datasets
2. More methods included in comparative analysis
3. New clustering method based on modified inversion formula density estimation

Authors

M.Lukauskas[1,2], T.Ruzgas[1]

1) Department of Applied Mathematics, Faculty of Mathematics and Natural Sciences, Kaunas University of Technology
2) Zyro Inc.