DATA CLUSTERING WITH MIXTURES OF MULTIDIMENSIONAL DISTRIBUTIONS

ABSTRACT

Unsupervised clustering is an important area in data analysis, machine learning, and artificial intelligence. The objective of this PhD project was to develop, implement, and compare model-based and distance-based unsupervised clustering algorithms, and evaluate their performance on various datasets using different metrics. The study aimed to address the challenge faced by data scientists when choosing the appropriate unsupervised clustering algorithm, given the many available methods, often accompanied by software implementations.

The study implemented two model-based algorithms, Gaussian Mixture EM and Multinomial Mixture EM, and compared it to four distance-based algorithms, agglomerative hierarchical clustering, k-means, k-medoids, and fuzzy c-means. The algorithms were applied to both simulated and actual datasets, and several metrics were used to quantify the clustering results, including Adjusted Rand Index, Simple Matching Coefficient, Weighted Jaccard Index, Balanced Accuracy, and metrics based on Beta-Binomial conjugate distribution. The findings showed that the model-based algorithms, particularly Gaussian Mixture EM and Multinomial Mixture EM, outperformed distance-based algorithms in many cases.

The study's contribution to the field of data analysis and machine learning was to provide insight into the development of more accurate and effective unsupervised clustering methods. The study demonstrated that model-based algorithms are potent tools in unsupervised clustering methods and are highly competitive compared to distance-based algorithms. Furthermore, the algorithms were implemented in R and made available on the GitHub platform.

The study conducted an extensive simulation study using thousands of Multivariate Gaussian Mixtures and Multinomial Mixtures to test the algorithms' performance. Additionally, a curated set of real datasets from various publicly available sources, including genomics/medical data, was also compared. Based on these datasets, hundreds of different components were prepared, with each group combination occurring only once in the same set, enabling a controlled comparison of the algorithms with differing numbers of parameters, dimensions, and clusters.

In summary, the study's results demonstrated that the model-based algorithms, particularly Gaussian Mixture EM and Multinomial Mixture EM, outperformed distance-based algorithms in many cases. The study contributes to advancing the field of data analysis and machine learning by informing the development of more accurate and effective unsupervised clustering methods. The developed algorithms are available in R and can be used in various applications, contributing to solving challenges in different scientific areas.