ISSN 2395-1621

A Survey On Clustering Algorithms

[#1]Kaushalya D. Korake, [#2]Prof. Vilas S. Gaikwad

[1]kaushalyakorake21@gmail.com
[2]vilasgaikwad11@gmail.com

[#1]Department of Computer Engineering
[#2]Prof. Department of Computer Engineering

JSPM Narhe Technical Campus,
Savitribai Phule Pune University, Pune
India.

ABSTRACT

In this work we presented the work related to the clustering using Euclidian distance and mahalanobis distance. The traditional methods used the Euclidian distance for the measuring similarity between the objects from dataset. This approach does not offer the good results in case of increased dimensionality. In this paper we also discussed the mahalanobis distance method to find the better clustering results. This paper also presents the work related to the basic types of clustering and their sub categories that have been used for clustering.

Keywords: Fuzzy-c-mean, K-means, Mahalanobis distance.

ARTICLE INFO

Article History

Received: 4th January 2016

Received in revised form :

6th January 2016

Accepted: 7th January, 2016

Published online :

11th January, 2016

I. INTRODUCTION

Clustering and classification are fundamental tasks in Data Mining. Classification is used mostly as a supervised learning method and Clustering forum supervised learning. The goal of clustering is descriptive that of classification is predictive [1].Clustering groups data items into subsets in such a manner that similar items are grouped together and different items belong to different groups [2]. The data instances are thereby organized into an efficient representation that characterizes the population being sampled. The main reason for having number of clustering approaches is the fact that the notion of‟ cluster" is not precisely defined. Farley and Raftery (1998) suggest dividing the clustering approaches into two main parts : hierarchical and partitioning approach. Hanand Kamber (2001) proposed categorizing the methods into additional two main categories [5]: density-based methods, model-based clustering.

A. Hierarchical Methods:

Hierarchical method construct clusters by partitioning clusters in either a top-down or bottom-up approach. This approach can be sub-categorized as follows [4]:

1. Agglomerative hierarchical clustering

In this approach each object initially represents cluster of its own. After that clusters are successively merged until the desired cluster structure is obtained.

2. Divisive hierarchical clustering

All objects initially belong to a single cluster. Then the cluster is divided into sub-clusters, again these clusters are successively divided into their own sub-clusters until the desired cluster structure is obtained.

B. Partitioning Methods

The partitioning approach relocates data items by moving from one cluster to another cluster. Starting from an initial partitioning, Such methods typically require the number of cluster to created. To achieve global optimality in partitioned-based clustering, an exhaustive enumeration process of all possible partitions is required [4].

1. **Error Minimization Algorithms.**
2. **Graph-Theoretic Clustering.**

C. Density-based Methods

Density-based methods assume that the points that belong to each cluster are drawn from a specific probability distribution. The overall distribution of the data is assumed to be a mixture of several distributions [1].The aim of these methods is to identify the clusters and their distribution parameters. These methods are designed for discovering clusters of arbitrary shape which are not necessarily convex [5].

D. Model-based Clustering Methods

These methods attempt to optimize the fit between the given data and some mathematical models. Unlike conventional clustering. Which identifies groups of objects, model-based clustering methods also find characteristic descriptions for each group, where each group represents a concept or class [3].

1. **Decision Trees.**
2. **Neural Networks**

II. LITERATURE SURVEY

1. Fuzzy and Crisp Recursive Profiling of Online Reviewers and Businesses:

This paper describes crisp and fuzzy meta clustering techniques to evolve two recursively defined clustering' schemes of both businesses and reviewers in parallel using areal-world dataset supplied by yelp.com [1]. The objective of paper is to profile the businesses and reviewers by grouping them based on similar characteristics. The novelty of the proposed approach is in the fact that the representations of both businesses and reviewers change dynamically throughout the meta clustering process [1]. A business is represented by static information obtained from the database and dynamic information obtained from the clustering of reviewers who reviewed the business. In the same way the reviewer representation augments the static representation from the database with profiles of businesses who are reviewed by these reviewers. The resulting web-based service provides a facility for users to find similar businesses/reviewers based on the category of the business, rating, number of reviews, and number of check-ins [1]. The paper provides a profile of a business or reviewer the factors with which users can put the reviews in context. Since an object can belong to multiple clusters in fuzzy meta clustering which makes it possible to absorb some of the groups consisting of outliers in one of the mainstream clusters [1].

2. A Clustering Algorithm based on Feature Weighting Fuzzy Compactness and Separation:

This paper aims at improving the well-known fuzzy compactness and separation algorithm (FCS). Paper proposes a new clustering algorithm based on feature weighting fuzzy compactness and separation (WFCS). The proposed algorithm introduces the feature weighting into the objective function [2]. System formulates the membership and feature weighting and analyses the membership of data points falling on the crisp boundary. The proposed WFCS is validated both on simulated dataset and real dataset [2]. The experimental results demonstrate that the proposed WFCS has the characteristics of hard clustering and fuzzy clustering and outperforms many existing clustering algorithms with respect to three metrics: Rand Index, Xie-Beni Index and Within-Between(WB) Index [2] .

The number of clustering algorithms are based on Euclidean distance as measure of similarity between data objects [2]. Such algorithms also require initial setting of parameters as a prior such as number of clusters. The Euclidean distance is very sensitive to scales of variables involved and independent of correlated variables [2]. To overcome drawbacks a hybrid clustering algorithm based on Mahalanobis distance is proposed in this paper [2]. The reason for the hybridization is to relieve the user from setting the parameters in advance.

3. Hybrid Clustering Algorithm based on Mahalanobis Distance and MST

In this paper we consider the performance of the widely adopted K-means clustering algorithm when the classification variables are correlated. In this paper authors measure performance in terms of recovery of the true data structure [3]. Performance worsens considerably if the groups have elliptical instead of spherical shape [3].Paper suggest some modifications to the standard K-means algorithm which can improve cluster recovery. Presented approach is based on a combination of careful seed selection techniques and use of Mahalanobis instead of Euclidean distances. Paper concludes that use of the Mahalanobis distance should become a standard option of the available K-means routines for non-hierarchical cluster analysis. This goal can be achieved by minor modifications in popular commercial software [3].

4. Mahalanobis clustering, with applications to AVO classification and seismic reservoir parameter estimation:

A clustering algorithm based on Mahalanobis distance is proposed as an improvement on traditional *K*-means clustering. Authors present applications of this method to both AVO classification and seismic reservoir parameter estimation using multiple attributes [5]. Latter application uses the radial basis function neural network (RBFN) with centres. After that it apply Mahalanobis clustering to find the cluster centres that are used in the training of the network [4]. Authors show that this method allows us to

improve the estimate of the covariance matrix parameters used in the general form of the RBFN approach [5].

5. Normalized clustering algorithm based on mahalanobis distance

FCM (fuzzy c-means algorithm) based on Euclidean distance function converges to a local minimum of the objective function. This can only be used to detect spherical structural clusters [5]. The added fuzzy covariance matrices in their distance measure were not directly derived from the objective function [6]. In this paper, an improved Normalized Clustering Algorithm Based on Mahalanobis distance by taking a new threshold value and a new convergent process is proposed [6].

6. Fuzzy C-Means Algorithm Based on Standard Mahalanobis Distances

Clustering technique plays an important role in data analysis and interpretation [7]. Fuzzy clustering is a branch in clustering analysis and it is widely used in the pattern recognition field [2]. Fuzzy clustering algorithms can only be used to detect the data classes with the same super spherical shapes [5].

GK-algorithm is a modified Mahalanobis distance with preserved volume based clustering approach. The added fuzzy covariance matrices in their distance measure are derived from the objective function directly [7].In GG algorithm Gaussian distance can only be used for the data with multivariate normal distribution [5].Some of the well-known fuzzy clustering algorithms are based on Euclidean distance function. Gustafson-Kessel clustering algorithm and Gath-Geva clustering algorithm were developed to detect non-spherical structural clusters. The improved Fuzzy C-Means algorithm based on different Mahalanobis distance, called FCM-M and FCM-CM were proposed [7].

III.CONCLUSION

This paper presented the study related to the different clustering approaches. Paper also presents the approaches related to the measuring the similarity between the objects. The paper presents the effectiveness of the mahalanobis distance method for large dimensional dataset.

REFERENCES

[1] Pawan Lingras And Matt Triff," Fuzzy And Crisp Recursive Profiling Of Online Reviewers And Businesses", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 23, NO. 4, AUGUST 2015.

[2] Yuan Zhou , Hong-Fuzuo 1 And Jiao Feng ,"A Clustering Algorithm Based On Feature Weighting Fuzzy Compactness And Separation", Www.Mdpi.Com/Journal/Algorithms2015, 8, 128-143.

[3] V. Vallikumari, BHVS Ramakrishnam Raju ," Hybrid Clustering Algorithm Based On Mahalanobis Distance And MST", International Journal Of Applied Information Systems (IJAIS) , Volume 3– No.5, July 2012.

[4] Andrea Cerioli,"K-Means Cluster Analysis And Mahalanobis Metrics: A Problematic Match Or An Overlooked Opportunity?",Statistica Applicata Vol. 17, N. 1, 2005.

[5] Brian H. Russell And Laurence R. Lines,"Mahalanobis Clustering, With Applications To AVO Classification And Seismic Reservoir Parameter Estimation ", CREWES Research Report — Volume 15 (2013).

[6] Jeng-Ming Yih ,Yuan-Horng Lin ," Normalized Clustering Algorithm Based On Mahalanobis Distance", International Journal Of Technical Research And Applications,Volume-2, Special Issue 2 (July-Aug 2014), PP. 48-52.

[7] Hsiang-Chuan Liu, Bai-Cheng Jeng, Jeng-Ming Yih, And Yen-Kueiyu,"Fuzzy C-Means Algorithm Based On Standard Mahalanobis Distances" Proceedings Of The 2009 International Symposium On Information Processing, Huangshan, P. R. China, August 21-23, 2009, Pp. 422-427.

