International Journal of Scientific Research and Management (IJSRM)
||Volume||5||Issue||08||Pages||6771-6776||2017||
Website: www.ijsrm.in ISSN (e): 2321-3418
Index Copernicus value (2015): 57.47 DOI: 10.18535/ijsrm/v5i8.26

Optimized Map Reduce Based Shuffling Mechanism for Density Clustering

Priya[1], Rakesh Chwala[2]

[1]M.Tech Scholar, CSE Department, Delhi Institute of Technology & Management, Sonipat
[2]Assistant Professor, CSE Department, Delhi Institute of Technology & Management, Sonipat

Abstract
Most international organizations produce more information in a week than many people could read in a lifetime. Everyday hundreds of megabytes of data are distributed around the world, but it is no longer possible to monitor this increasingly rapid development – the growth is nearly exponential. So the basic problems with the management of the data are its high Dimensionality. Dimensionality reduction can also be seen as the process of deriving a set of degrees of freedom which can be used to reproduce most of the unpredictability. This work is about Dimensionality Reduction, which in turn is about converting data of very high dimensionality into data of much lower dimensionality such that each of the lower dimensions conveys much more information.

Keywords: KDD , DBMS , OLAP , STC

I. Introduction

Data mining discovers description through clustering visualization, association, and sequential analysis. Clustering is a primary data description method in data mining which groups' most similar data. Data clustering is a common technique for data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. The work in this thesis is related to clustering algorithm. So it becomes important to have an overview of the concept of clustering.

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. Clustering techniques fall into a group of undirected data mining tools. The goal of undirected data mining is to discover structure in the data as a whole. Clustering techniques are used for combining observed objects into clusters (groups), which satisfy two main criteria: Each group or cluster is homogeneous; objects that belong to the same group are similar to each other. Each group or cluster should be different from other clusters, that is, objects that belong to one cluster should be different from the objects of other clusters. Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects, which are "similar" between them and are "dissimilar" to the objects belonging to other clusters

The organization of present paper is as follow. Section II presents the literature survey which highlights the facts of various researchers. Section III describes the methodology used for proposed work as in this paper ant colony optimization is used. Result analysis is presented in section IV following the concluding remarks in section V.

II. Literature review

This section will provide the brief description and highlights the contribution, remarks and factors of the work done by the researchers. Many attempts have been made in the past to achieve minimization of mean square error & execution time.

Jain, et. al, in 2010[L1] describes Organizing data into sensible groupings is one of the most fundamental modes of understanding and learning. The aim of clustering is to find structure in data and is therefore exploratory in nature. Clustering has a long and rich history in a variety of scientific fields.

Qinbao Song, et al , in “A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data” 2013, the authers propose a feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. The results, on 35 publicly available real-world high dimensional image, microarray, and text data, demonstrate that FAST not only produces smaller subsets of features In this paper, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data.

Shi Yu , et al in “Optimized Data Fusion for Kernel k-Means Clustering” presents a novel optimized kernel k-means algorithm (OKKC) to combine multiple data sources for clustering analysis. The algorithm uses an alternating minimization framework to optimize the cluster membership and kernel coefficients as a non-convex problem. In the proposed algorithm, the problem to optimize the cluster membership and the problem to optimize the kernel coefficients are all based on the same Rayleigh quotient objective; therefore the proposed algorithm converges locally.

Venkateswaran K., et al in “Change Detection in Synthetic Aperture Radar Images Using Contourlet Based Fusion and Kernel K-Means Clustering” presents change detection algorithms play a vital role in overseeing the transformations on the earth surface. In kernel K-means clustering, non-linear clustering is performed, as a result the false alarm rate is reduced and accuracy of the clustering process is enhanced. The aggregation of image fusion and kernel K-means clustering is seen to be more effective in detecting the changes than its preexistences.

Dharmendra K Roy and Lokesh K Sharma in presented Clustering is one of the major data mining tasks and aims at grouping the data objects into meaningful classes (clusters) such that the similarity of objects within clusters is maximized, and the similarity of objects from different clusters is minimized. In this work, we have proposed a modified genetic k-means algorithm for finding a globally optimal partition of a given mixed numeric and categorical data into a specified number of clusters.

Christos Boutsidis , et al in “Randomized Dimensionality Reduction for k-means Clustering” presented Dimensionality reduction encompasses the union of two approaches: *feature selection* and *feature extraction*. A feature selection based algorithm for k-means clustering selects a small subset of the input featur then applies k-means clustering on the selected features. A feature extraction based algorithm for k-means constructs a small set of new artificial features and then applies k-means clustering on the constructed features.

Singaravelu.S, A.Sherin and S.Savitha in “Agglomerative Fuzzy K-Means Clustering Algorithm” present an agglomerative fuzzy K-Means clustering algorithm for numerical data, an extension to the standard fuzzy K-Means algorithm by introducing a penalty term to the objective function to make the clustering process not sensitive to the initial cluster centers. In this paper, we have presented a new approach, called the agglomerative fuzzy K-Means clustering algorithm for numerical data to determine the number of clusters. The new approach minimizes the objective function, which is the sum of the objective function of the fuzzy k-mean and the entropy function. Our experimental results have shown the effectiveness of the proposed algorithm when different initial cluster centers were used and overlapping clusters are contained with data sets.

V.S.V.S. MURTHY et al in “Content Based Image Retrieval using Hierarchical and K-Means Clustering Techniques” present an image retrieval system that takes an image as the input query and retrieves images based on image content. Content Based Image Retrieval is an approach for retrieving semantically-relevant images from an image database based on automatically-derived image features. Hierarchical clustering assists faster image retrieval and also allows the search for most relevant images in large Thus using hierarchical and K-Means techniques together not only facilitates the user not to overlook the image he may require but also to obtain accurate favored image results.

This section has provided the brief review of the work done in past. It also highlighted the factors, contribution and remarks on the achievement.

III. Frame Work for Implementation

The main objectives of research work are, To Collect High Dimensional data for analysis, Clustering high-dimensional data is the cluster analysis of data with anywhere from a few dozen to many thousands of dimensions.

1. Randomly select *‘c’* cluster centers.
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4. Recalculate the new cluster center using:

$$V_i = \left(\frac{1}{C_i}\right) \sum_{j=1}^{Ci} X_i$$

5. where, ‘c_i’ represents the number of data points in i^{th} cluster.
6. Recalculate the distance between each data point and new obtained cluster centers.
7. If no data point was reassigned then stop, otherwise repeat from step 3).
8. Initialize *dim* array = size(data)
9. For each row successfully clustered using K-means,
10. Find Cluster c to which the row r belong
11. Dim[i] =c; //the c cluster center can now represent all the data points
12. // A, B, C .. with single value c
13. Return Dim

There’s another way to deal with clustering problems: a model-based approach, which consists in using certain models for clusters and attempting to optimize the fit between the data and the model.

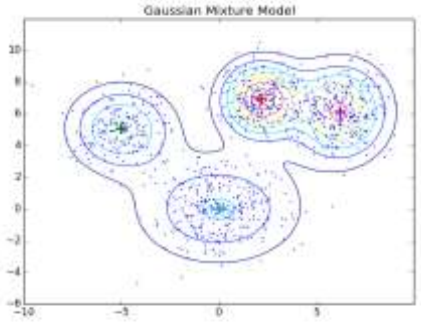


Fig 1 Example of Gaussian Mixture Model

In practice, each cluster can be mathematically represented by a parametric distribution, like a Gaussian (continuous) or a Poisson (discrete). The entire data set is therefore modelled by a mixture of these distributions. An individual distribution used to model a specific cluster is often referred to as a component distribution. A mixture model with high likelihood tends to have the following traits: component distributions have high “peaks” (data in one cluster are tight); the mixture model “covers” the data well (dominant patterns in the data are captured by component distributions).

Result Analysis

MATLAB platform has been used to evaluate the results. Some assumptions are made to simulate the results as discussed The different parameters used in proposed work are given in table 1. The performance parameters are analyzed using MATLAB 2016a.

For dimensionality reduction various datasets have been taken for application, One such dataset is defined below, where only two dimensions exists specified in two columns namely

A: Height of student(inches)

B: Weight of Student (pounds)

The dataset taken is normalized and realistic in nature, This data collection was designed to ascertain the nutritional status of students. These data supply information on students' age, year of birth, height, and weight, state and county of birth, and state and county of residence.

Table 1 Student Height and Weight Dataset

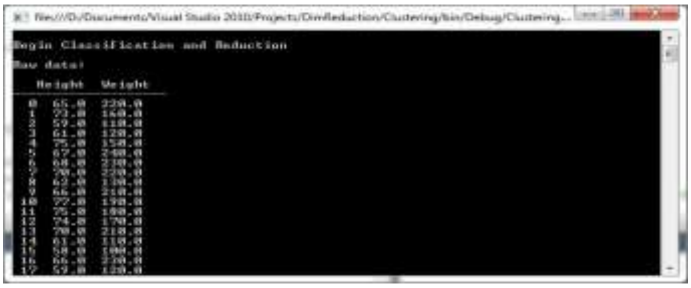
Height	Weight
65	220
73	160
59	110
61	120
75	150
67	240
68	230
70	220
62	130
66	210
77	190
75	180
74	170
70	210
61	110
58	100
66	230
59	120
68	210
61	130

Table 2 Classified Student Height and Weight Dataset

Height	Weight	Cluster ID
65	220	1
73	160	3
59	110	0
61	120	0
75	150	3
67	240	1
68	230	2
70	220	2
62	130	0
66	210	1
77	190	3
75	180	3
74	170	3
70	210	2
61	110	0
58	100	0
66	230	1
59	120	0
68	210	2
61	130	0

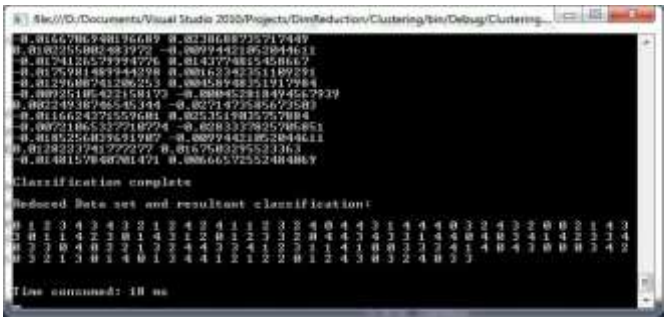
Reduces Dimension of data can be elaborated in following way

1	3	0	0	3	1	2	2	0	1	3	3	3	2	0	0	1	0	2	0



```
Begin Classification and Reduction
Raw data:
    Height   Weight
  ---------------------
   0   65.0   220.0
   1   73.0   160.0
   2   59.0   110.0
   3   61.0   120.0
   4   75.0   150.0
   5   67.0   240.0
   6   68.0   230.0
   7   70.0   220.0
   8   62.0   130.0
   9   66.0   210.0
  10   77.0   190.0
  11   75.0   180.0
  12   74.0   170.0
  13   70.0   210.0
  14   61.0   110.0
  15   58.0   100.0
  16   66.0   230.0
  17   59.0   120.0
```

Fig .2 Importing Data Set From CSV file



```
-0.0166706940196689 0.0238688735717449
0.0102255002483972 -0.00994421052044611
-0.0174126579994776 0.0143774815458667
-0.0175981489944298 0.00162342351109291
-0.0129600741206253 0.000458948351917984
-0.00925105422158173 -0.000452818494567939
0.00224938706545344 -0.0271473585673580
-0.0116624271559601 0.0253519035757804
-0.00721065327710774 -0.0283337025706851
-0.0185256039691987 -0.00994421052044611
0.0128223741277277 0.0167503295523363
-0.0148157040701471 0.00666572552484069

Classification complete

Reduced Data set and resultant classification:

0 1 2 3 4 3 4 3 2 1 2 4 2 4 1 1 2 3 2 4 0 4 4 3 1 4 4 4 0 3 2 4 3 2 0 0 2 1 4 3
3 0 1 1 4 2 3 0 1 4 3 1 2 0 1 2 3 3 2 0 4 4 3 4 3 3 1 4 4 0 4 0 3 4 1 4 2 3 3 4
0 2 2 0 4 0 2 2 1 2 2 4 4 2 2 4 1 2 2 1 1 4 1 0 0 3 3 2 4 1 4 0 4 3 0 0 0 3 4 2
0 3 2 1 3 0 1 4 0 1 3 4 4 1 2 1 2 2 0 1 2 4 3 0 3 2 4 0 3 3

Time consumed: 18 ms
```

Fig 3 Reduced data set & time Consumed

Table.3 Classification of 150 records of different dataset with actual classes

	Student	Fisher's iris	Image Segmentation	Glass identification
Time	15	15	18	36
Classes	2	4	5	7

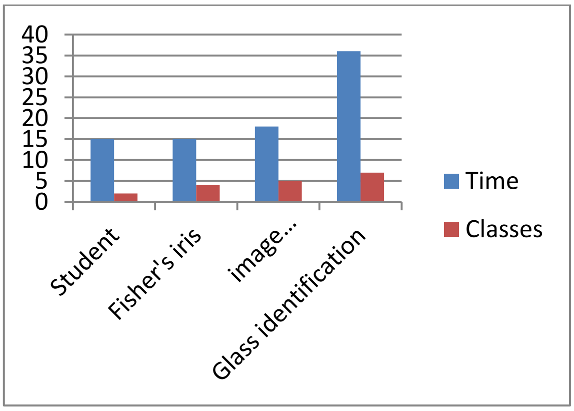


Figure 5 Time vs Classes of different dataset with actual classes

IV. Conclusion

Dimensionality reduction can also be seen as the process of deriving a set of degrees of freedom which can be used to reproduce most of the unpredictability. Dimensionality Reduction is about converting data of very high dimensionality into data of much lower dimensionality such that each of the lower dimensions conveys much more information.

This research work was about reducing N-Dimensional datasets, this work adopted the infamous K-means the algorithm for the Dimensionality reduction of datasets. The algorithm can successfully reduce any numeric high dimensional dataset thus reducing the complexity of underlying data. The Algorithm can be repeatedly applied to multidimensional dataset to further reduce its dimensions

References:

1. Beil, F.; Ester, M. and Xu, X. (2005), “Frequent Term-Based Text Clustering”, Germany SIGKDD 05 Edmonton, Alberta, Canada Copyright 2005 ACM 1-58113-567-X/05/0007.
2. Berkhin, P. (2002). “Survey of Clustering Data Mining Techniques”, Technical report, Accrue Software, San Jose, CA, 2002.
3. Cheung, Y.M. (2003), “k-Means: A new generalized k-means clustering algorithm”, Pattern Recognition Letters 24 (2003) 2883–2893, 2003.
4. Cui, X. and Potok, T.E. (2005), “Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm”, Applied Software Engineering Research Group, Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6085, USACuix, 2005.
5. Fahim, A.M.; Salem, A.M.; Torkey, F.A. and Ramadan, M.A. (2006), “An efficient enhanced k-means clustering algorithm”, Journal of Zhejiang University SCIENCE

A, ISSN 1009-3095 (Print); ISSN 1862-1775, 2006.

6. Guha, S.; Meyerson, A.; Mishra, N.; Motwani, R. and Callaghan, L. (2003), “Clustering Data Streams: Theory and Practice”, IEEE transactions on Knowledge and Data Engineering, Vol. 15, No. 3, May/June 2003.
7. Han, H.; Manavoglu, E.; Giles, L. and Zha, H. (2003), “Rule-based Word Clustering for Text Classification”, SIGIR’03, July 28–August 1, 2003, Toronto, Canada, ACM 1581136463/03/0007, 2003.
8. Hatzivassiloglou, V.; Klavans, J.L. and Eskin, E. (1999), “Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning”, Center for Research on Information Access, Columbia University, New York, 1999.
9. Hotho, A.; Staab, S. and Stumme, G. (2003), “Text Clustering Based on Background Knowledge”, Institute of Applied Informatics and Formal Description Methods AIFB, University of Karlsruhe, D–76128 Karlsruhe, Germany, Technical Report N0.425, September 2003.
10. Jagannathan, G.; Pillaipakkamnatt, K. and Wright, R.N. (2006), “A New Privacy-Preserving Distributed k-Clustering Algorithm”, Proceedings of the 2006 SIAM International Conference on Data Mining (SDM), 2006.
11. Jain, A.K.; Murty, M.N. and Flynn, I.J. (2004), “Data Clustering: A Review”, ACM Computing Surveys, Vol. 31, No. 3, September 1999.
12. Kanungo, T.; Mount, D.M.; Netanyahu, N.S; Piatko, C.D; Silverman, R. and Wu, A.Y. (2002), “An Efficient k-Means Clustering Algorithm: Analysis and Implementation”, IEEE Transactions of Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, July 2002.
13. Karypis, G.; Han, E.H. and Kumar, V. (1999), “CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling”, IEEE Computer, 32(8): pp. 68-75, 1999.
14. Law, M.H.C.; Topchy, A.P. and Jain, A.K. (2004), “Multiobjective Data Clustering”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004.
15. Maitra, R.; Peterson, A.D. and Ghosh, A.P. (2011), “A systematic evaluation of different methods for initializing the k-means clustering algorithm”, IEEE Transactions on Knowledge and Data Engineering, Vol. 10, No. , pp.23-238, 2011.
16. Martinez, I.O.A.; Trinidad, J.F.M. and Ochoa, J.A.C. (2005), “Conceptual K-Means Algorithm with Similarity Functions”, Springer-Verlag Berlin Heidelberg 2005, CIARP 2005, LNCS 3773, pp. 368 – 376, 2005.