International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol.2, Issue.4, July-Aug. 2012 pp-1482-1485 ISSN: 2249-6645

Robust Information Clustering on MDI

Mr. S. M. Krishna Ganesh[1], Dr. S. Venkatesan Jaya Kumar[2]
[1](Department of Computer Science and Engineering, St.Joseph College of Engineering and Technology, India)
[2] (Department of General Engineering St.Joseph College of Engineering and Technology, India

Abstract : We propose a robust framework for determining a natural clustering of a given dataset, based on the minimum description length (MDL) principle. The proposed framework, robust information-theoretic clustering (RIC), is orthogonal to any known clustering algorithm, Given a preliminary clustering, RIC purifies these clusters from noise, and adjusts the clustering's such that it simultaneously determines the most natural amount and shape of the clusters. RIC, for refining a clustering and discovering a most natural clustering of a dataset. In particular, we propose a novel criterion, volume after compression (VAC), for determining the goodness of a cluster, and propose algorithms for robust estimation of the correlation structure of a cluster in the presence of noise, identification and separation of noise using VAC, and construction of natural correlation clusters by a merging procedure guided by VAC. It is fully automatic, that is, no difficult or sensitive parameters must be selected by the user. Our RIC method can be combined with any clustering technique ranging from K- means and K-medoids to advanced methods such as spectral clustering. The various performance related graphs shows that our algorithm is computationally and performance wise better than other clustering algorithms.

Keywords - Clustering techniques, Minimum description Length, Volume after compression

I. INTRODUCTION

In General, Clustering mean the assignment of objects into groups so that objects from the same cluster are more similar to each other than objects from different clusters.

Clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. 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. There is no target variable to be predicted, thus no distinction is being made between independent and dependent variables. Depending on the clustering technique, clusters can be expressed in different ways identified clusters may be exclusive, so that any example belongs to only one cluster. They may be overlapping; an example may belong to several clusters. they may be probabilistic, whereby an example belongs to each cluster with a certain probability. clusters might have hierarchical structure, having crude division of examples at highest level of hierarchy, which is then refined to sub- clusters at lower levels. The problem of clustering has attracted a huge volume of attention for several Decades, with multiple books (Hartigan,Van-Rijsbergen), surveys (Murtagh) and papers (X- means (Pelleg and Moore), G-means (Hamerly and Elkan), CLARANS (Ng andHan),CURE (Guha), CLIQUE (Agrawal), BIRCH(Zhang),DBSCAN(Ester).Recent interestin clustering has been on finding clusters that have non-Gaussian correlations in subspaces of the attributes, for example, the work of Bohm,Tung and Aggarwal and Yu. Finding correlation clusters has diverse applications ranging from spatial databases to bioinformatics. The hard part of clustering is to decide what is a good group of clusters, and which data points to label as outliers and thus ignore for clustering. We show a fictitious set of points in 2D. Shows a grouping of points that most humans would agree is g od a Gaussian like Cluster at the left, a line-like cluster at the right, and a few noise points (outliers) scattered throughout. However, typical clustering algorithms like How can we quantify the goodness of a grouping? We would like a function that will give a good score to the grouping of and a bad one to that of diagram. How can we write an algorithm that will produce good groupings efficiently and without getting distracted by outliers..

II. MOTIVATION

Clustering has attracted a huge volume of interest. Recently, there have been several papers focusing on scalable clustering algorithms, such as, CURE[Guha et al. 1998], CLIQUE [Agrawal et al. 1998], BIRCH [Zhang et al. 1996], DBSCAN [Ester et al. 1996], and OPTICS [Ankerst et al. 1999]. There are also parameter- free algorithms like X-means [Pelleg and Moore 2000], and G-means [Hamerly and Elkan 2003]. However, they all suffer from one or more of the following drawbacks: They focus on spherical or Gaussian clusters, and/or are sensitive to outliers, and/or need user-defined thresholds and parameters. Most algorithms are geared towards Gaussian or plain spherical clusters; for example, the well-known K- means algorithm, BIRCH[Zhang et al.1996] (which is suitable for spherical clusters), X-means [Pelleg and Moore 2000], and G- means [Hamerly and Elkan 2003]. These algorithms tend to be sensitive to outliers because they try to optimize the log-likelihood of a Gaussian, which is equivalent to the Euclidean (or Mahalanobis) distance— either way, an outlier has high impact on the clustering.

Density-based clustering methods, such as DBSCAN and OPTICS, can detect clusters of arbitrary shape and data distribution and are robust against noise. For DBSCAN the user has to select a density threshold, and for OPTICS to derive clusters from the reachability plot. K-harmonic means [Zhang et al. 2000] avoids the problem of outliers, but still needs *k*. Spectral

clustering algorithms [Ng et al. 2001] perform K-means or similar algorithms after decomposing the $n \times n$ gram matrix of the data (typically using PCA). Clusters of arbitrary shape in the original space correspond to Gaussian clusters in the transformed space. Here also k needs to be selected by the user. Recent interest in clustering has been on finding clusters that have non-Gaussian correlations in subspaces of the attributes [B¨ohm et al. 2004; Tung et al. 2005; Aggarwal and Yu 2000]. Finding correlation clusters has diverse applications ranging from spatial databases to bioinformatics. the information bottleneck method [Tishby et al. 2000], which is used by Slonim and Tishby for clustering terms and documents [Slonim and Tishby 2000], and the work of Still and Bialek [2004]. Based on information theory they derivea suitable distance function for coclustering, but the number of clusters still needs to be specified in advance by the user.Clustering is important for many applications such as Library, city planning. There are several motivations for robust clustering, but the basic requirement is to remove the noise .Previously used techniques does not able to cluster the group in efficient manner because does not contains sufficient information to cluster, cannot able to activate in fast manner, cannot able to group the given data is efficient manner and computations are performed is a tough manner.

III. PROPOSED ALGORITHM RIT

The following illustrates the proposed algorithm for Robust Information Theoretical Clustering

Input: Dataset as input

Output: To Identified & analysis the best pair with minimal cost as output.

Step1: Give the input dataset.
Step2: The input dataset is filtered by using noise removal method, then purify the cluster dataset by removing noise.
Step3: After purifying, merge the dataset.
Step4: Find the best pair of clusters to merge.
Step 5: Identify and analyze the best pair with minimal cost of dataset.

IV. PURIFICATION OF NOISE

The first step of purifying a cluster of points is to identify the proper decorrelation matrix. We generate several estimates (called candidates) of the covariance matrix, using various estimation methods, and pick the one with the best overall VAC value. In our experiments, the candidates include the matrix $\_C$ from the conventional method using arithmetic average, and matrix $\_R$ from the robust method described earlier. We also determine a conventional and a robust candidate, matrices $\_C$, 50 and $\_R$, 50, respectively, by considering only a certain percentage (e.g., 50%) of points in the cluster being closest to the robustly estimated center $\_ìR$. In addition,we always have the identity matrix I as one candidate decorrelation matrix. Among these matrices, our algorithm selects the matrix giving the best (lowest) overall VAC.

The next step is to detect noise points in the cluster. By now, we have computed the robust center $\_ìR$, and have chosen a candidate covariance matrix which we call $\_*$ (the corresponding decorrelation matrix is $V*$). The goal is to partition the set of points in cluster C into two new sets: *Ccore* (for core points) and *Cout* (outliers). First, our method orders the points of C according to the Mahalanobis distance defined by the candidate covariance matrix $\_*$. Initially, we define all points to be outliers (*Cout* = *C*, *Ccore* = {}). Then, we iteratively remove points $\_x$ from *Cout* (according to Mahalonobis sort order starting with the point closest to the center) and insert them into *Ccore*, and compute the coding costs before and after moving the point $\_x$.

V. MERGING FROM NOISE

Our RIC framework is designed to refine the result of any clustering algorithm.Due to imperfection of the clusters given by an algorithm, our cluster purifying algorithm may lead to redundant clusters containing noise objects that fit well to other neighboring noise clusters. Our algorithm corrects the wrong partitions by merging clusters that sha re common characteristics, taking into account the subspace orientation and data distribution. We use the proposed VAC value to evaluate how well two clusters fit together.

The idea is to check whether the merging of a pair of clusters could decrease the corresponding VAC values. Our proposed merging process is an iterative procedure. Our algorithm merges those two clusters which have the maximum *savedCost* (.) value, resulting in a greedy search toward a clustering that has the minimum overall cost. To deter this greedy algorithm from getting stuck in a local minimum, we do not stop immediately, even when no savings of *savedCost* (., .) value can be achieved by merging pairs of clusters. In other words, we do not stop when *savedCost*(., .) = 0. Instead, the algorithm continues for another t iterations, continuing to merge cluster pairs (Ci , Cj) with the maximum *savedCost*(Ci , Cj) value, even though now the *savedCost* (Ci , Cj) value is negative and merging Ci and Cj will increase the VAC value of the overall dataset. Whenever a new minimum is reached the counter is reset to zero.

VI. EXPERIMENTAL RESULTS

We proposed to compare the noise cluster, purified cluster and merged cluster with Optimization values and maximal cost values. When compared purified cluster is better than other clusters. In the figure 1,We compared noise cluster, purified cluster and emerging cluster with optimization values. Then we Compare best pair of merging cluster with purified cluster.

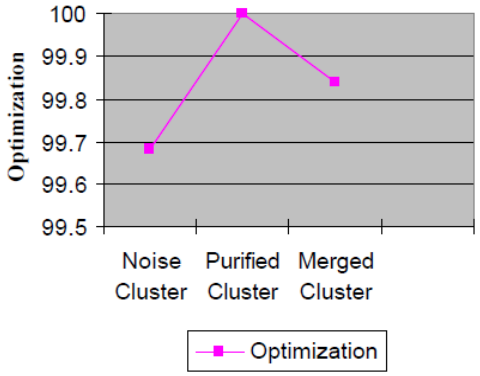

Figure 1. Compare best pair of merging cluster

In the figure 2, we compared noise cluster, purified cluster and merging cluster with maximal cost. values. Compare best pair of merging cluster with the purified cluster

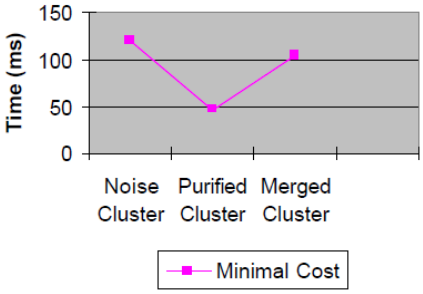

Figure 2. Compare the best pair of merging cluster with maximal cost

Type	Optimization	Maximal Cost (ms)
Noise Cluster	99.68531	120
Purified Cluster	100	47
Merged Cluster	99.840	104

Table: Values of Optimization and Maximal cost in Noise cluster, purifies cluster and merged clusters

VII. CONCLUSION

In this paper, RIC framework is designed.We have input dataset filtered to those data by using noise removal method,then purify clusters dataset by removing noise. After data set TyA is purified merge with the dataset. Find the best pair of clusters to merge. To identify and analysis the best pair with minimal cost of dataset. Previous work clustering methods, such as DBSCAN and OPTICS, can detect clusters of arbitrary shape and data distribution and are robust against noise. For DBSCAN the user has to select a density threshold, and for OPTICS o derive clusters from the reachability plot. K-harmonic means avoids the problem of outliers, but still needs *k*. There are several motivations for robust clustering, but the basic requirement is to remove the noise .Previously used techniques does not able to cluster the group in efficient manner because it does not contains sufficient information to cluster, unable to activate in fast manner and unable to group the given data in efficient manner and computations are performed in this method

VIII. ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their insightful remarks and valuable suggestions, which were very helpful in improving the article. We thank our respected Rev.Fr.Dr.Arulraj Founder, DMI Group of institutions, East Africa and India, Dr.T.X.A.Ananth, Director, Mr. Ignatius Herman, Director (Academic),Dr.V.Krishnan Ph.D, Principal, DMI.St.Joseph College of Engg & Technology, Tanzania. for their encouragement and support. Also we thank our parents, friends and colleagues for their support and encouragement.

REFERENCES

[1] C.C.Aggarwal and P. S. Yu, "Finding generalized projected clusters in high dimensional spaces", In *SIGMOD Conference*, 2000, pp. 70–81.

[2] Frank Rehm1, Frank Klawonn2 and Rudolf Kruse "*A Novel Approach to Noise Clustering for Outlier Detection*" , IEEE communication , 2007, pp.1-5.

[3] Dave RN, Krishnapuram R (1997), "*Robust clustering methods a unified view*", IEEE Transactions magazine, pp.270-293.

[4] M. Ester, H.P. Kriegel, J.Sander, and X. Xu "A density-based algorithm for discovering clusters in large spatial databases with noise". in *KDD Conference*,1996.

[5] S. Guha, R. Rastogi, and K. Shim , "CURE: An efficient clustering algorithm for large databases", in *SIGMOD Conference*, pp.73–84,1998.

[6] Rajesh N. Davé and Sumit Sen, *"Robust Fuzzy Clustering of Relational Data"*, IEEE Transaction Magazine,pp.713-727,2002

[7] Aggarawal C. C. and Yu P. S, "Finding generalized projected clusters in high dimensionalspaces", *Proceedings of the ACM International SIGMOD Conference on Management of Data*, pp.70–81, 2000.

[8] Agrawal R., Gehrke J, Gunopulos, D, and Raghavan. P , " Automatic subspace clustering of high dimensional data for data mining applications", *in Proceedings of the ACM International SIGMOD Conference on Management of Data ,pp. 94–105,1998*.

[9] Bhattacharya, V. Ljosa, J.-Y. Pan, M. R. Verardo,H. Yang, C. Faloutsos, and A. K. Singh, " ViVo: Visual vocabulary construction for mining biomedical images",*IEEE International Conference on Data Mining (ICDM),,2005*

[10] Ankerst . M., Breunig. M, Kriegel . H.-P., and Sander.J, "OPTICS: Ordering points toidentify the clustering structure", *in Proceedings of the ACM International SIGMOD Conference on Management of Data,1999* .

[11] Banfield J. D. and Raftery A. E , " *Model-based Gaussian and non-Gaussian clustering*, Biometrics 1993,pp.803–821.

[12] Bohm C.,Kailing. K.,Kroger. P., and Zimek A., "Computing clusters of correlation connected Object", *in Proceedings of the ACM International SIGMOD Conference on Management of Data, pp.455–466,*2004.

[13] Chakrabarti D., Padimitriou, S., Modha D. S., and Faloutsos C. "Fully automatic cross associations",*in Proceedings of the ACM SIGKDD Conference on International Knowledge Discoveryand Data Mining,2004,.pp. 79–88.*

[14] Guha S., Rastogi, R., and Shim. K (1998), "CURE: An efficient clustering algorithm for large databases",*in Proceedings of the ACM International SIGMOD Conference on Management of Data,pp. 73–84,1998*.

[15] Hamerly, G. and Elkan, C. ., "Learning the k in k-means", *in Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS),2003.*

[16] Murtagh. F, "A survey of recent advances in hierarchical clustering algorithms", *in Proceedings of the ACM SIGKDD Conference on International Knowledge Discovery and Data Mining. pp. 354–359,1983.*

[17] Christian Bohm,Christos Faloutsos,Claudia Plant , "Outlier- robust clustering using independent components", i*n Proceedings of the ACM International SIGMOD Conference on Management of Data,2000.*

[18] Alexander Hinneburg and Daniel A. Keim, *"An efficient approach to clustering in large multimedia databases with noise", in Knowledge Discovery and Data Mining, pp . 58 – 65,1998.*

[19] Olfa Nasraoui,Carlos Rojas,"Robust Clustering for Tracking Noisy Evolving Data Streams", *in Proceedings of the ACM international SIGMOD Conference on Management of Data,1999.*

[20] Jia Liao ,Bo Zhang, "A Robust Clustering Algorithm for video shot using Haar Wavelet Transformation", *:in Proceedings of SIGMOD, pp. 730- 741,2007.*

[21] Eike achtert,Christian Bohm,Jorn David,Peer Kroger,Arthur Zimek ,"Robust Clustering in Arbitrarily Oriented Subspaces", *in Proceedings of SIGMOD, pp. 124- 126,2000.*

[22] Anupam Joshi,Raghu Krishnapuram, "Robust Fuzzy Clustering Method to support web mining" , *in proceedings of the ACM International SIGMOD Conference on Management of Data,pp.400 – 407,2000.*