ISSN: 2321-7782 (Online)
Volume 2, Issue 11, November 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Attribute and Information Gain based Feature Selection Technique for Cluster Ensemble: Hybrid Majority Voting based Variable Importance Measure

K N V Ramya Devi[1]
M.Tech Scholar, Cse Department
Lenora College Of Engineering
Rampachodavaram, India

M Srinivasa Rao[2]
Asst.Prof., Cse Department
Lenora College Of Engineering
Rampachodavaram, India

Abstract: ***Cluster analysis is one of the prominent unsupervised learning techniques widely used to categorize the data items based on their similarity. Mainly off-line and online analysis through clusters is more attractive area of research. But, high dimensional big data analysis is always introducing a new dimension in the area of data mining. Because high dimensional cluster analysis is giving less accurate results and high processing time when considering maximum dimensions. To overcome these issues dimensionality reduction techniques have been introduced. Here, a million dollar questions are, which dimensions are to be considered? , what type of measures have to be introduced? And how to evaluate the cluster quality based on those dimensions and measures? In this paper, we are trying to propose a novel hybrid technique for dimensionality reduction for better cluster analysis. Proposed algorithm will be completed in polynomial time.***

Keywords: Feature Selection, Cluster Ensemble, Information Gain, Attribute Clustering

I. INTRODUCTION

The intention of cluster analysis is to form set of groups of data objects based on some similarity or dissimilarity measures among those objects. Any similarity measure tries to identify the level of coincidence of values of properties of objects. Based on those observations objects are divided into multiple groups such that every group epitomizes unique behaviour.

The challenge of high dimensional data clustering is materialized due to swift extension of our competence in automatic generation of data and possession. Among available properties or attributes of real time entities only some of them are significant requirement. Such properties are called as relevant variables and rest are called as irrelevant. Apart from the existing problems in cluster analysis such as choosing of optimal number of clusters, computation complexities and absence of decisive factor for cluster quality and result analysis. Further, irrelevant variables' presence could make the cluster as outlier. The results of the algorithms those depends on sub space and pair wise similarity are affected by these irrelevant variables.

Dimensionality reduction or variable selection is a common approach as a data pre-processing step of most of the cluster analysis techniques. The aim of this technique is to detect most relevant variables to form efficient sub spaces. This technique has been approached from various perspectives: model-based methods, density based methods and criterion based methods.

Literature Survey:

Density based variable selection methods:

CLIQUE proposed by Agrawal et al. (1998) scans each variable individually and removes those variables with no high dimension region. Remaining variables are carried out for sub space clustering. One lagging point of CLIQUE is that it doesn't consider the separation among the regions. Chan and Hall (2010) proposed statistical based analysis for feature selection.

Model based variable selection methods:

Overview of the model based methods was presented in Mclachlan and Peel, 2000.Three methods have been proposed in model based clustering for variable selection. Fully Bayesian approached has been proposed by Tadesse et al. (2005). Raftery and Dean (2006) proposed another method which indulgence the selection of variable problem as model comparison questions' sequence and use Bayesian Information Criteria to compare 2 nested models. Pan and Shen (2007) proposed likelihood and log likelihood methods. By finding the maximum log likelihood variable selection has been performed in this method.

Criterion-based variable selection methods:

Example of these methods includes K-Means. These methods use some condition to be satisfied from previous result of iteration to current iteration's result. K-Means uses sum-of-squares to measure the change in clusters from iteration to iteration.

As in Witten and Tibshirani (2010), the BCSS is defined as

$$\begin{aligned} BCSS &= \sum_{j=1}^{a} BCSS(\boldsymbol{X}_{\cdot j}, \boldsymbol{G}) \\ &\equiv \sum_{j=1}^{d} \left(\frac{1}{2n} \sum_{i=1}^{n} \sum_{i'=1}^{n} (x_{ij} - x_{i'j})^2 - \sum_{k=1}^{K} \frac{1}{2N_k} \sum_{i,i' \in G(k)} (x_{ij} - x_{i'j})^2 \right), \end{aligned}$$

where X.j is the j-th column of the data matrix X, G(k) denotes k-th group, Nk is the size of the k-th group, and xij the i-th observation of the j-th variable. The quantity BCSS (X.j ;G) measures the between cluster sum of squares of the j-th variable when the group labels are assigned as in G.

Witten and Tibshirani (2010) proposed a sparse K-Means method by maximizing the following criterion

$$\begin{aligned} \text{maximize}_{\boldsymbol{G},\mathbf{w}} \quad & \sum_{j=1}^{d} w_j \, BCSS(\boldsymbol{X}_{\cdot j}, \boldsymbol{G}) \\ & \text{subject to } ||\mathbf{w}||_{L_2}^2 \leq 1, \; ||\mathbf{w}||_{L_1} \leq s, \; w_j \geq 0 \quad \forall j. \end{aligned}$$

with maximization over both clustering assignment G and the weight vector w. Following algorithm explain this.

Algorithm 1 Iterative Algorithm of Sparse K-Means

Set $w_1 = \cdots = w_d = \frac{1}{\sqrt{d}}$
repeat
  Perform weighted K-Means on X with $w_1, \cdots, w_d$ to get a clustering G
  Calculate $BCSS(\boldsymbol{X}_{\cdot j}, \boldsymbol{G})$ for $j = 1, \cdots, d$
  Set $w_j = 0$, if $BCSS(\boldsymbol{X}_{\cdot j}, \boldsymbol{G}) \leq \delta$
    $w_j = BCSS(\boldsymbol{X}_{\cdot j}, \boldsymbol{G}) - \delta$, if $BCSS(\boldsymbol{X}_{\cdot j}, \boldsymbol{G}) > \delta$
until convergence
Variables with non-zeros weights are the selected relevant variables.

II. Proposed Algorithm

This paper is proposing a novel hybrid method of variable selection. According to this method data set is partitioned based on the individual attribute sub spaces. Later clustering is applied on this ensemble of partitions. Then evaluate the quality of each cluster. Now arrange the clusters in decreasing order based on quality because every cluster is single attribute cluster. So quality is indirectly specifies the significant of the variable up to what extent it is relevant to cluster. Now input the threshold or upper limit of number of dimensions to be considered among the available attribute clusters.

Second stage of process includes calculation of information gain for each attribute. Here also sort the attributes in descending order based on information gain. Now select the attributes those are common to both attribute clustering results and

information gain process. Thus these variables or attributes are treated as most relevant variables to do clustering. Do clustering based on these variables and evaluate cluster quality. This entire process is described in the following algorithms

III. ALGORITHMS

Algorithm-1: Attribute wise Clustering (AC)

Input: Dataset, similarity threshold, max dimension limit

Output: relevant attribute set (RA={A1,A2,A3...})

1. Read the dataset header part
2. Split the dataset or vertically partition the dataset equal to number of available attributes
3. Now apply incremental clustering technique on this ensemble with given similarity threshold.
4. Evaluate the cluster quality of each clustering result.
5. Sort the attribute cluster results in descending order
6. Select top most attributes within given max dimension limit as relevant variables

Algorithm-2: Information Gain based Variable Selection(IGVS)

Input: Dataset, max dimension limit

Output: relevant attribute set (RA={A1,A2,A3...})

1. Let DT as giving dataset and assume that every instance of the form(X,y)={x1,x2,...,xn,y) where X= set of attributes {xi} and yi= class label of current instance Xi. xi belongs to vals(Xi) of ith attribute.
2. Information gain for an attribute a_i is defined in terms of entropy H() as follows

$$IG(T,a) = H(T) - \sum_{v \in vals(a)} \frac{|\{\mathbf{x} \in T | x_a = v\}|}{|T|} \cdot H(\{\mathbf{x} \in T | x_a = v\})$$

3. Repeat the above step for all the available attributes
4. Sort the attributes in descending order based on the information gain
5. Select the attributes by following the criterion top most attributes <= max dimension limit

The following algorithm is used to follow the Attribute wise clustering and Information Gain for variable selection as a pre-processing step for link based similarity of cluster ensemble.

Algorithm-3: Link based similarity of cluster ensemble with ACIGVS

Input: Dataset, Similarity Threshold, max dim limit, final clustering threshold

Output: Set of Clusters and their quality

1. Let ACA be the set of relevant attributes extracted from the Attribute Clustering
2. i.e ACA=AC(dataset,simth,max-dim-limit)
3. Let IGA be the set of relevant attributes extracted from the Information Gain
4. i.e IGA=IGVS(dataset,max-dim-limit)
5. Let RA be the final relevant attribute set

6. i.e RA=ACA ∩ IGA

7. Apply Link based similarity on cluster ensemble with relevant attribute set (RA)

Problem Formulation and General Framework

Let X ={ x1,…xN} be a set of N data points and ∏={∏1,… ∏M} be a cluster ensemble with M base clusterings, each of which is referred to as an ensemble member. Each base clustering returns a set of clusters $\pi_i = \{C_1^i, C_2^i, \ldots, C_{k_i}^i\}$. For each x ε X, C(x) denotes the cluster label to which the data point x belongs. The problem is to find a new partition ∏*of a data set X that summarizes the information from the cluster ensemble ∏.

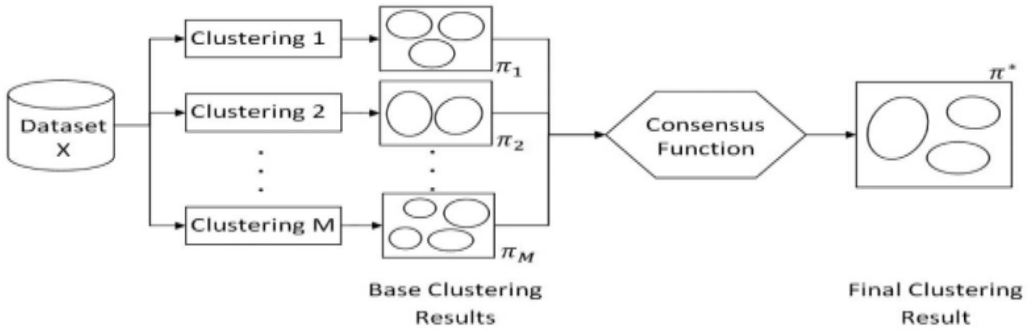


Figure 1 Cluster Ensemble Architecture

The above figure shows the general framework of cluster ensembles.

A Dataset X is taken and it is given to some base clustering algorithms and the solutions achieved from different base clustering are aggregated to form a final partition. This meta-level methodology involves two major tasks of:

1. Generating a cluster ensemble with relevant attribute extracted from step 6 of algorithm-3

2. Producing the final partition normally referred to as a consensus function.

IV. ENSEMBLE GENERATION METHODS

For data clustering, the results obtained with any single algorithm over much iteration are usually very similar. In such a circumstance where all ensemble members agree on how a data set should be partitioned, aggregating the base clustering results will show no improvement over any of the constituent members. As a result, several heuristics have been proposed to introduce artificial instabilities in clustering algorithms, giving diversity within a cluster ensemble. The following ensemble generation methods yield different clustering of the same data, by exploiting different cluster models and different data partitions.

1. Homogeneous ensembles: Base clustering's are created using repeated runs of a single clustering algorithm, with several sets of parameter initializations, such as cluster centers of the k-means clustering technique.

2. Random-k: One of the most successful techniques is randomly selecting the number of clusters (k) for each ensemble member.

3. Data subspace/sampling: A cluster ensemble can also be achieved by generating base clustering's from different subsets of initial data. It is intuitively assumed that each clustering algorithm will provide different levels of performance for different partitions of a data set.

4. Heterogeneous ensembles: A number of different clustering algorithms are used together to generate base clustering's.

5. Mixed heuristics: In addition to using one of the aforementioned methods, any combination of them can be applied as well.

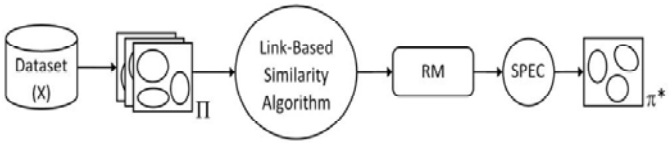
In this project we are using Homogeneous increment cluster ensemble technique.

Consensus Functions: Having obtained the cluster ensemble, a variety of consensus functions have been developed and made available for deriving the ultimate data partition. Each consensus function utilizes a specific form of information matrix, which summarizes the base clustering results.

Link based Cluster Ensemble Approach

It includes three major steps

1. Creating base clustering's to form a cluster ensemble (∏)
2. Generating a refined cluster-association matrix (RM) using a link-based similarity algorithm
3. Producing the final data partition (∏*) by exploiting the spectral graph partitioning technique as a consensus function.



Where RM is Refinement Matrix and is defined as follows

$$RM(x_i, cl) = \begin{cases} 1, & if cl = C_*^t(x_i), \\ sim(cl, C_*^t(x_i)), & otherwise, \end{cases}$$

Where xi be the current instance, cl be the class label and $C^t(x_i)$ is a cluster label (corresponding to a particular cluster of the clustering ∏t) to which data point xi belongs and Sim(Cx,Cy) ε [0,1] denotes the similarity between any two clusters C x, Cy, which can be discovered using the following link-based algorithm.

This paper followed "*Subspace ensemble*", a method to generate diversity within an ensemble is to exploit a number of different data subsets. To this extent, the cluster ensemble is established on various data subspaces, from which base clustering results are generated.

V. EXAMPLE AND DISCUSSION

Language: Java

Dataset: Temporal and Categorical Baseball Dataset

Operating System: Window/Linux

Processor: Pentium IV or later

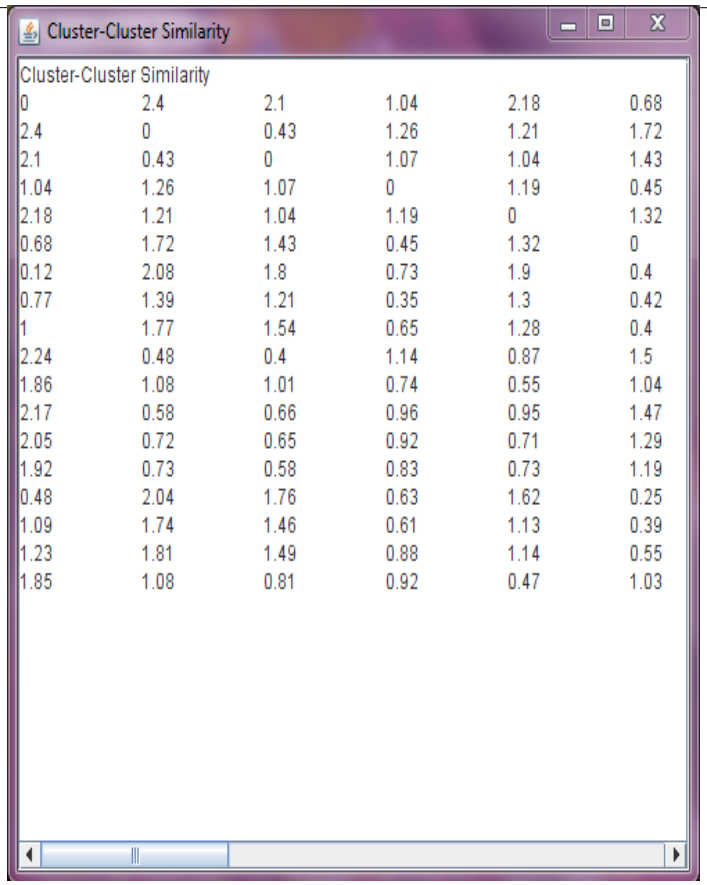


Figure 2 Refinement Matrix

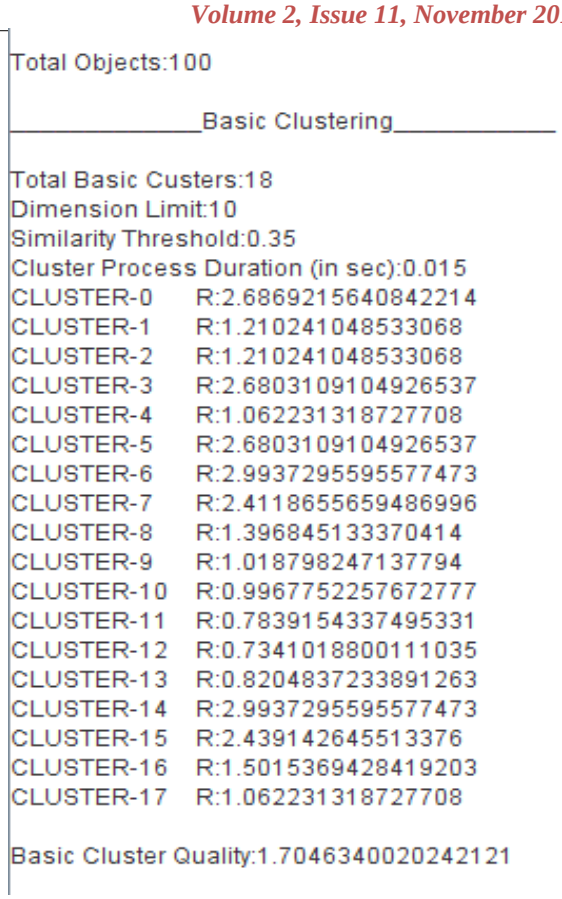


Figure 3 First Clustering Results

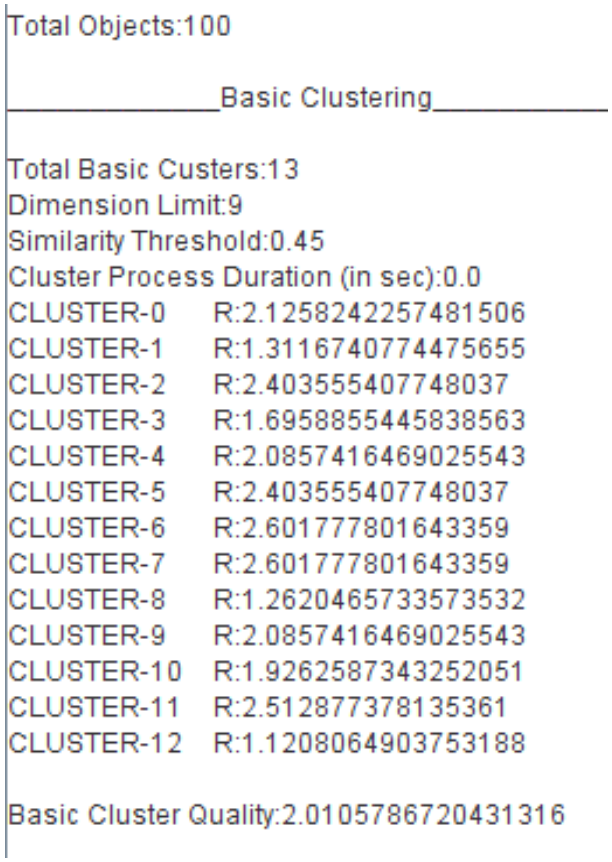


Figure 4 Second Clustering Results

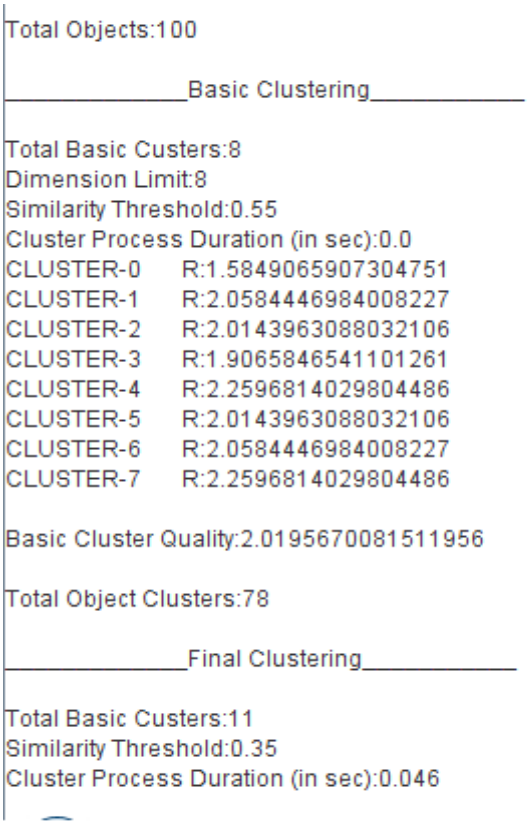


Figure 5 Third Clustering in Ensemble and Final Clustering Results

VI. Conclusion and Future Work

In this paper, we are trying to propose a novel hybrid technique for dimensionality reduction for better cluster analysis. Proposed algorithm will be completed in polynomial time. Simulation results shows each cluster quality, total objects placed in each cluster, process duration and final cluster results applied on a temporal and categorical dataset. Here we need to consider the type of clustering algorithm for more optimum results and auto parameter tuning techniques to achieve better cluster quality. This will be our future work.

References

1. J.C. Gower, "A General Coefficient of Similarity and Some of Its Properties," Biometrics, vol. 27, pp. 857-871, 1971.
2. V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS: Clustering Categorical Data Using Summaries," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 73-83, 1999.
3. Y. Yang, S. Guan, and J. You, "CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 682- 687, 2002.
4. S. Monti, P. Tamayo, J.P. Mesirov, and T.R. Golub, "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data," Machine Learning, vol. 52, nos. 1/2, pp. 91-118, 2003.
5. X.Z. Fern and C.E. Brodley, "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning," Proc. Int'l Conf. Machine Learning (ICML), pp. 36-43, 2004.
6. N. Iam-On, T. Boongoen, and S. Garrett, "Refining Pairwise Similarity Matrix for Cluster Ensemble Problem with Cluster Relations," Proc. Int'l Conf. Discovery Science, pp. 222-233, 2008.

Author(s) Profile

K N V RAMYA DEVI, received the B.Tech degree in Computer Science and Engineering in jntu,KAKINADA in 2011, worked as teaching Assistant in engineering college for 2 years

M SRINIVAS working as Asst.Professor in cse dept of Lenora engineering college since 2009,completed his M,Tech in computer science engineering from ANU Guntur, studied B.Tech in computer science engineering from JNTU kakinada

