International Journal of Engineering & Technology, 7 (2.11) (2018) 27-30

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Feature Selection using Genetic Algorithm for Clustering high Dimensional Data

Kahkashan Kouser[1*], Amrita Priyam[2]

[1,2]Birla Institute of Technology, Ranchi, India.
**Corresponding author E-mail: kahkashankouser@gmail.com*

Abstract

One of the open problems of modern data mining is clustering high dimensional data. For this in the paper a new technique called GA-HDClustering is proposed, which works in two steps. First a GA-based feature selection algorithm is designed to determine the optimal feature subset; an optimal feature subset is consisting of important features of the entire data set next, a K-means algorithm is applied using the optimal feature subset to find the clusters. On the other hand, traditional K-means algorithm is applied on the full dimensional feature space. Finally, the result of GA-HDClustering is compared with the traditional clustering algorithm. For comparison different validity matrices such as Sum of squared error (SSE), Within Group average distance (WGAD), Between group distance (BGD), Davies-Bouldin index(DBI), are used .The GA-HDClustering uses genetic algorithm for searching an effective feature subspace in a large feature space. This large feature space is made of all dimensions of the data set. The experiment performed on the standard data set revealed that the GA-HDClustering is superior to traditional clustering algorithm.

Keywords: *feature selection; clustering; high dimensional data; Genetic algorithm.*

1. Introduction

Clustering is a descriptive method, which assign a collection of data element into subset called clusters, such that the data elements in the same cluster are more similar than the data element of another cluster [1]. There are many new clustering algorithm has been developed in the past years, which perform the clustering of low dimensional data in a very grateful manner. When we are going to apply these traditional clustering algorithms to the high dimensional data set, they do not perform well. There are many factors which are responsible for it. First high dimensional data set contains irrelevant and redundant dimensions, which completely mask the objective clusters, secondly curse of dimensionality; it is concerned with the sparsity of data. In higher dimensional space, the data objects turn into very sparse [2].

So dimension reduction methods are used for reducing the number of dimensions for clustering high dimensional data set. Feature selection and feature transformation are two commonly used methods for dimension reduction.

Feature transformation is a preprocessing step, which permitting the clustering algorithm to use simple few of newly created feature. Some of clustering algorithm has integrated such feature transformation technique to identify essential feature and iteratively enhance their clustering result, but these techniques don't really eliminate or withdraw any feature from original feature space [3]. When there is a large member of irrelevant attributes which hide the clusters, these techniques appear worthless because these irrelevant features are still present in the data set, which hide the clusters. Another drawback of using a combination of attribute is, they are very hard to interpret.

Feature selection is a very useful technique for removing irrelevant and redundant attribute of high dimensional feature space. For a given set of high dimensional feature space it will find the subset of features which might be most appropriate for the data mixing task [4].
A feature selection procedure consists of four major steps as shown below [5]:

- Subset creation
- Assessment of subset
- Stopping conditions
- Result validation

The subset creation procedure creates a new feature subset. This newly created featured subset is evaluated based on some evaluation criteria. This new featured subset is compared with the previous best feature subset based on some evaluation criteria. If the new feature subset is better, it will replace the previous best feature subset. The procedure of subset creation and evolution is continued until stopping criteria is satisfied. Finally the chosen feature subset is commonly validated via prior knowledge or different test through artificial and/or real-world data sets.
In this paper, we proposed a genetic algorithm based feature selection method, which use the searching capability of genetic algorithm to find a suitable feature subspace for clustering high dimensional data called GA-HD clustering.

In the next section a description of genetic algorithm and its application is given in section 2. In the section 3, a detail description of the GA-HDclustering is given. An experiment is done on the seed data set to show its capability and efficiency. In section 4 the result

of an experiment is discussed and comparison of performance of the GA-HD clustering is done. Finally, section 5 contains the conclusion.

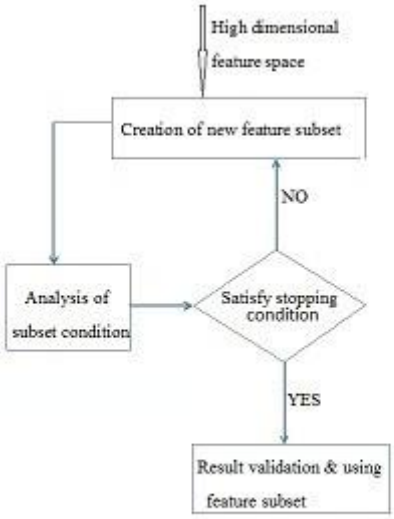

Fig. 1. Process Flow

2. Related Work

The process of genetic algorithm includes three important stages. Initialization of population, fitness calculation of the population, and the creation of new population [6].

As a first new population is created randomly which consists of a member of individual (or chromosome).Then a fitness value is assigned to every individual which is computed by using a fitness function or objective function. Fitness function measures the quality of the individual. Higher fitness function means better individually. Creation of new population is done with the help of three genetic operators' selection, crossover and mutation. The process of applying selection, crossover and mutation are repeated for a constant number of times until stopping criteria is not satisfied.

A brief introduction of Different cluster validation measure, which is used for measuring goodness of clustering result, is given below.

2.1. Sum of squared error(SSE)

Sum of squared error (SSE) is calculated by summing the squared distance between the cluster centroid (Cen_i) and every object within the cluster [7]. To determine the best cluster configuration we try to minimize SSE.

$$SSE = \sum_{i=1}^{k} \sum_{o \epsilon Ci} d(Ceni, O)^2 \quad (1)$$

Here K represent the number of cluster ,Cen_i represent centroid of the *ith* cluster and *O* represent object belonging to a cluster.

2.2 WGAD-BGD

Within Group average distance (WGAD) is a tool for measuring the cohesion, and Between group distance (BGD) is a tool for measuring separation. WGAD is obtained by summing over the average distance between the cluster centroid i and every object of the cluster [8].

$$\text{Total WGAD} = \sum_{i=1}^{i=k} \frac{\sum_{j=1}^{j=mi} dist(xj,Ceni)}{mi} \quad (2)$$

Where *k* represents the number of clusters and ***may*** represent the number of objects in a cluster. *xj* represent the object, and Cen_i represent the centroid of the cluster.
BGD is calculated by summing over the distance between each cluster centroid Cen_i and overall cluster centroid C.

$$\text{Total BGD} = \sum_{i=1}^{k} dist(Ceni, C) \quad (3)$$

The difference between pair of WGAD and BGD is used to express the cluster validity. Thus to determine a best cluster configuration we try to minimize WGAD-BGD measure.

2.3 Davies-Bouldin Index (DBI)

Davies-Bouldin index is used to measure similarity between the clusters (***Rij).*** $\mathbf{R_{ij}}$ is calculated by

$$R_{ij} = \frac{Si - Sj}{dij} \quad (4)$$

Where d_{ij} is the distance between the centroid of cluster *i* and cluster *j*. $S_i(S_j)$ is a dispersion measure used to determine average distance between the objects belonging to the same cluster[8].

$$Si = \frac{\sum_{n=1}^{n=|ci|} d(xn,Ceni)}{|Ci|} \quad (5)$$

Here *Ceni* represent the centroid of cluster i , xn is used to represent the object within cluster. |Ci| represent the number of objects within a cluster.

The Davies-Bouldin index is calculated for every cluster pair. Finally, we add maximum cluster similarity of every cluster to make a single DB-index.

$$DB = \frac{\sum_{i=1}^{i=k} Ri}{k} \quad (6)$$

where $Ri = \max_{j=1,\, i \neq j} (R_{ij})$, i =1.....k
k is the number of cluster.

Low value of Davies-Bouldin index indicates that the clusters are not very similar, which means clusters are well-separated and compact.

3. High-Dimensional Data Clustering Using Genetic Algorithm

3.1. Individual Representation for Applying Genetic Algorithm and Creation of Initial Population

Generally two encoding methods are used for representing individual, binary encoding floating point encoding. Binary encoding use a search space which is larger than floating point encoding, but the crossover and mutation operation can perform more easily on it. So, we use binary encoding for this paper.
An individual solution of the solution space or search space consists of two parts (CR, CS), CR is binary string used to represent feature subspace and CF is used to represent the fitness value.
For example, if an individual has representation (1010101,0.7489), means the feature subspace is consist of 1st, 3rd ,4th & 7th attribute and have the fitness value 0.7489.
If the original feature set has N attribute, then the initial population have 2^n-1 individual.

3.2. Fitness Function

To determine how good an individual can be in solution space, a fitness function or objective function is applied to every individual in solution space. At each generation the fitness level

of the individual is calculated. The convergence and searching capability of GA is affected largely by the fitness function
The correlation coefficient between the attribute is used to compute the fitness value of each chromosome or individual in the population.
Correlations between the attributes are determined by the formula [9].

$$\text{CORRELATION(s)} = \frac{m\,C attribute-class}{\sqrt{(m+m(m-1)\beta C attribute-attribute}} \quad (7)$$

S=denotes the subset of attributes which are currently selected.
M=denotes the number of attribute in subsets. $C_{attribute\text{-}class}$=is average attribute –class correlation.
$C_{attribute\text{-}attribute}$=is an average correlation between attributes.
β=scaling factor (0.25)
Higher the value of correlation coefficient better is the quality of candidate solution with respect to the problem into consideration.

Table 1. Parameters Observed

	Class	**Area**	**Perimeter**	**Compactness**	**Length of Kernel**	**Width Of**	**Asymmetry Co-**	**Length of kernel**
Class	1	0.285	0.557	0.556	0.205	0.400	0.580	0.084
Area	0.28510	1	0.994376	0.623801	0.949234	0.961131	0.214584	0.859383
Perimeter	0.557699	0.994376	1	0.546374	0.971823	0.945443	0.197938	0.886877
Compactness	0.55699	0.623801	0.546374	1	0.385108	0.773071	0.367904	0.237481
Length of kernel	0.205198	0.949234	0.971823	0.385108	1	0.860197	0.146009	0.930147
Width of kernel	0.400023	0.961131	0.945443	0.773071	0.860197	1	0.257069	0.744318
Asymmetry	0.580581	0.214584	0.197938	0.367904	0.146009	0.257069	1	0.021035
Length of kernel	0.084585	0.859383	0.886877	0.237481	0.930147	0.744318	0.021035	1

3.3. Genetic Operator

The genetic operator plays very important role in genetic algorithms [10]-[14]. Genetic operator is applied to the current population to create new population for the next generation. Genetic operator is a powerful tool for controlling an evaluation process.

3.3.1 Selection: Selection is the first operation applied to the population. On the selection, based on the fitness value chromosome are chosen from the population, to be parent to cross over and produce new offspring. In this paper roulette wheel selection strategy is used for selecting an individual.
3.3.2 Crossover: If a chromosome is m. A bit longer than a crossover site is chosen from range [1, m-1]. Binary string from beginning to crossover side is chosen from one chromosome and the rest is chosen from another chromosome, to create a new chromosome.
3.3.3 Mutation: In the mutation operation some randomly selected bits are inverted. In GA- HD clustering mutation assigns a probability Pm for each bit of the chromosome to change from 0 to 1 or vice versa.
3.3.4 Termination Criteria: Genetic algorithm run over a number of generations until termination criteria is not satisfied. Some termination criteria are given below:

- Maximum number of generations is reached.
- When the fitness value of a chromosome reached to a specified value.

4. Implementation Results

To illustrate the capability and efficiency of the GA-HD clustering, we perform it on the seed data set and compare it with the conventional means algorithm. The seed data set consists of 200 instances and seven features. The number cluster is three. The distribution of instance to the class is shown below:
Class1 has 70 elements, class2 has 68 elements, Class3 has 60 elements.

The entire data set can be classified into 3 classes, but when we are performing clustering, the result is not very good so we perform feature selection to determine the optimal subset of features by using genetic algorithm. Single point Crossover operation with crossover rate 0.7is applied. The Mutation rate is 0.5 .The algorithm runs in 50 generations. The chromosome 1101001 has maximum fitness value 0.925360. So subset of 4 attributes, namely (area, Perimeter, Length of the kernel, Width of the kernel, Length of kernel groove) constructs an optimal feature subset. In 2nd stage this optimal feature subset is used to perform clustering.

Table 2: Experimental result for GA-HD clustering with selected attribute.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Total**
SSE	15.259918	19.130043	21.168098	55.558060
WGAD-BGD	0.798269	0.843199	0.779278	3.624872
DBI	0.007149	0.007149	0.007385	0.021682
No. of iterations				600

Table 3: Experimental result from conventional clustering with all dimensions.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Total**
SSE	201.488068	184.108551	158.870682	544.467285
WGAD-BGD	2.769699	2.715463	2.385417	7.615520
DBI	0.056057	0.006185	0.056057	0.056057
No. Of iteration				800

When we compare both algorithms on SSE, we find that the SSE value by using all dimensions of the data set have resulted as follows: Cluster 1 has SSE value 201.488068, Cluster 2 have 184.108551, Cluster 3 has 158.870682 Which cause overall SE value 544.467285, while the SSE value of GA-HD clustering is as follows: Cluster 1 have 15.259918, Cluster 2 have 19.130043, Cluster 3 have 21.168098.This results overall SSE value 55.558060 which is less than the value of traditional clustering algorithm.

When we compare on WGAD-BGD, we find out the traditional clustering have values as follows Cluster 1 :2.769699, Cluster 2 : 2.715463, Cluster 3:2.385417.This results in overall value 7.615520.While the GA-HD clustering have values: Cluster 1: 0.798269, Cluster 2: - 0.843199, Cluster 3 :0.779278. Which results in overall WGAD BGD value 3.624872 Which is lesser than the traditional algorithm.

Similarly, when we calculate the DBI it have overall value, 0.021682 for GA –HD clustering and 0.118299 for traditional clustering algorithm. Final comparison is based on the number of iteration, GA-HD clustering have total number of clustering 600

and traditional clustering algorithm have 800.On the basis of all these measures, we determine GA-HD clustering is better.

5. Conclusion

In this paper GA-HD clustering is described which use genetic algorithm for clustering high- dimensional data set. It finds an effective subspace by using GA. Correlation metric is used as fitness function. Higher the value of coefficient function higher is the subspace. It is determined from experiment that GA-HD Clustering is more effective for clustering high dimensional data set. Fitness function proposed that the paper is effective for clustering, high dimensional data set, but new fitness function can be designed which can further improve the result, so we can say that GA-HD clustering is effective and powerful. It can determine relatively better subspace.

References

[1] Sun, M., Xiong, L., Sun, H., & Jiang, D. (2009, October), A GA-based feature selection for high-dimensional data clustering. In 3rd International Conference on Genetic and Evolutionary Computing WGEC'09, pp. 769-772.

[2] Sun, H. J., & Xiong, L. H. (2009, August), Genetic algorithm-based high-dimensional data clustering technique. In Sixth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD'09, Vol. 1, pp. 485-489.

[3] Parsons, L., Haque, E., & Liu, H. (2004), Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explorations Newsletter* 6, 90-105.

[4] Alzubaidi, A., Cosma, G., Brown, D., & Pockley, A. G. (2016, October), Breast cancer diagnosis using a hybrid genetic algorithm for feature selection based on mutual information. In International Conference on Interactive Technologies and Games (iTAG), pp. 70-76.

[5] Tiwari, R., & Singh, M. P. (2010), Correlation-based attribute selection using genetic algorithm. *International Journal of Computer Applications* 4, 28-34.

[6] Li, J. (2015, December), A feature subset selection algorithm based on feature activity and improved GA. In 11th International Conference on Computational Intelligence and Security (CIS), pp. 206-210.

[7] Chaimontree, S., Atkinson, K., & Coenen, F. (2010, November). Best clustering configuration metrics: towards multiagent based clustering. In International Conference on Advanced Data Mining and Applications (pp. 48-59). Springer, Berlin, Heidelberg.

[8] David Bouldin Index, Available at: https://en.wikipedia.org/wiki/DavieBouldin_index

[9] Hall, M. A. (1999). Correlation-based feature selection for machine learning.

[10] Rostami, M., & Moradi, P. (2014, May), A clustering based genetic algorithm for feature selection. In 6th Conference on Information and Knowledge Technology (IKT), pp. 112-116.

[11] Desale, K. S., & Ade, R. (2015, January), Genetic algorithm based feature selection approach for effective intrusion detection system. In International Conference on Computer Communication and Informatics (ICCCI), pp. 1-6.

[12] Song, Q., Ni, J., & Wang, G. (2013), A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering* 25, 1-14.

[13] Chandrashekar, G., & Sahin, F. (2014), A survey on feature selection methods. *Computers & Electrical Engineering* 40, 16-28.

[14] Goldberg, D. E. (1989), Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley.

[15] Han, J., Pei, J., & Kamber, M. (2011), Data mining: concepts and techniques. Elsevier.

[16] Dunham, M. H. (2006), Data mining: Introductory and advanced topics. Pearson Education India..