Available online at www.elixirpublishers.com (Elixir International Journal)

Advanced Engineering Informatics

Elixir Adv. Engg. Info. 40 (2011) 5174-5178

ISSN: 2229-712X

Rough genetic approach of data clustering

Rajashree Dash, Rajib Lochan Paramguru and Rasmita Dash

Department of Computer Science and Engineering, Siksha O Anusandhan University, Bhubaneswar, Orissa, India.

ARTICLE INFO

Article history:

Received: 22 August 2011;

Received in revised form:

17 October 2011;

Accepted: 27 October 2011;

Keywords

Data clustering,
Attribute Reduction,
Rough Set Theory,
Genetic Algorithm.

ABSTRACT

Due to the development of new techniques for generating and collecting data, the rate of growth of scientific databases has become large, which creates both a need and an opportunity to extract implicit knowledge to analyze these datasets. Analysis of such large expression data gives rise to a number of new computational challenges not only due to the increase in no. of data objects but also due to the increase in no of attributes. Hence to improve the efficiency and accuracy of mining task on high dimensional data, the data must be preprocessed by an efficient dimensionality reduction method. In this paper, we have proposed a Rough Genetic Approach for high-dimensional data clustering. Initially an efficient method of Rough Set Theory has been applied on the discritized data set to generate a reduced set of relevant attributes. Then, it is proposed to use the Genetic Algorithm for finding the cluster index of the dataset with reduced attribute which may give better clustering accuracy than other clustering techniques.

Introduction

Data mining is a convenient way of knowledge extraction from large data sets and focusing on issues relating to their feasibility, usefulness, effectiveness and scalability. Clustering is one of the commonly used data mining task that subdivides an input data set into a desired number of subgroups so that members of the same subgroup will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation and also by its ability to discover some or all of the hidden patterns. Traditional clustering techniques do not provide better results for very high dimensional datasets. Hence, attribute reduction or dimensionality reduction is an essential data-preprocessing task for cluster analysis of datasets having a large no. of features/attributes.

Rough set theory (RST) is a commonly used attribute reduction method, which is applied as a tool to discover data dependencies and to reduce the no. of attributes contained in the data set using the data alone, requiring no additional information [1] ,[2] . Given a dataset with discretized attributes, it is possible to find a reduct of original attributes that are most predictive of the class attribute. Rough set reducts can be found by using degree of dependency or using discernibility matrix.

In this paper, it has been proposed initially to apply the Rough set theory to generate the reduced set of necessary attributes or to construct the core of the attribute set by finding the upper and lower approximation of the reduced data set. The main advantage of this approach stems from the fact that this framework is able to characterize the granulation structure of a rough set using a granulation order. Then, it is proposed to use the Genetic Algorithm for finding the cluster index of the dataset with reduced attributes. Genetic Algorithm (GA) is a technique used to find approximate solutions to search problems through application of the principles of evolutionary biology. Genetic Algorithms use biologically inspired techniques such as genetic inheritance, natural selection, mutation, and sexual reproduction (recombination, or crossover) [3].

The GA consists of an iterative process that evolves a working set of individuals called a population toward an objective function, or fitness function. Traditionally, solutions are represented using fixed length strings, especially binary strings, but alternative encodings have been developed. Before applying the steps of GA to clustering, we propose to find the suitable value of maximum number of possible clusters for the dataset. Then the centers of the clusters may be placed anywhere in the space. Then the solutions will be genetically kept modified .The greatest advantage of genetic algorithms is that we can alter the fitness function to change the behavior of the algorithm [4].

Related Work

A novel approach to clustering using Genetic Algorithm has proposed in [5]. Here, Genetic Algorithm is mainly used for the purpose of controlled nature of clustering with a modified fitness function. Another improved genetic algorithm method has been proposed in [6]. In this paper all individuals are encoded by floating-point number and the sum of mean square deviation of intra-class distance is adopted as the objective function. By following the steps of the Genetic algorithm, the experimental results showed that the accuracy of GA can reach over 98 percent and generated better clustering results than other clustering techniques.

A genetic algorithm-based high-dimensional data clustering technique, called GA An improved Quick reduct Algorithm to select the features from the information system has been proposed in [8]. In this paper the Quick reduct algorithm has implemented for medical database of UCI machine learning repository and over the real HIV data set which consists of both numeric and non-numeric attributes. A new rough set based feature selection approach called Parameterized Average Support Heuristic (PASH) that considers the overall quality of the potential set of rules and selects features causing high

Tele:
E-mail addresses: rajashree_dash@yahoo.co.in

average support of rules over all decision classes, has proposed in [9].

Proposed Model

Rough Genetic approach of data clustering is a combination of feature selection method with the optimization technique like genetic algorithm for finding the cluster index of high dimensional continuous data set. The work of this paper can be implemented by collecting continuous data sets from data repository and by applying the degree of dependency approach of rough set theory to obtain a set of discriminative features. The steps of genetic algorithm can now be applied on the reduced dataset to obtain more accurate cluster index. As the rough set theory can be applied only on discretized data set, hence a suitable unsupervised discretization technique has proposed to apply on the original continuous data set. Then the decision table needs to be created for RST. A suitable approach of finding maximum no. of possible cluster centers has also applied before the steps of genetic algorithm to find the cluster index of each data set.

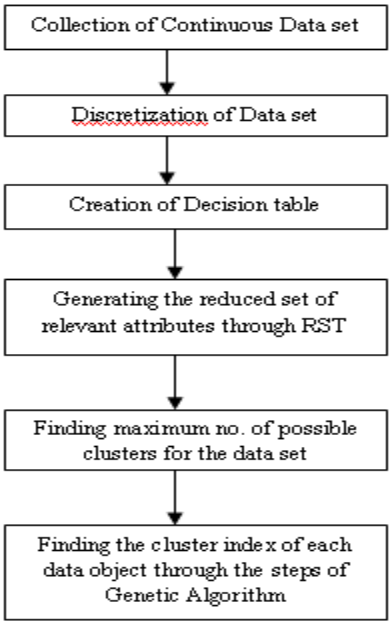


Figure 1 Proposed Rough Genetic Approach of Data Clustering

Experimental Analysis and Result Discussion

The proposed algorithm has evaluated on a synthetic dataset with 10 data objects having 4 attributes as shown in table 1. Initially the structure of rows has used to find the subsets of features that are highly correlated and represent each group optimally in terms of high spread in the lower dimension, reconstruction and insensitivity to noise. The steps of the algorithm are given below.

Step 1: Discretization of Dataset

The continuous dataset needs to be discretized before the application of RST. Comparisons of various discretization techniques are proposed in [10], [11]. Here we preferred to use discretization using k-means clustering as it uses minimum square error partitioning to generate an arbitrary number k of partitions reflecting the original distribution of the partition attribute and also it can perform equally well to that of supervised methods [12]. The output of discretization has shown in table 2.

Step 2: Creation of Decision Table

A decision table containing object ids, the discretized attributes and the decision attribute is created. The class attribute of the data set are normally used as the decision attribute. For unsupervised datasets the cluster index obtained can be used as the decision attribute. For the synthetic dataset, the cluster index obtained by applying the K-Means algorithm to the whole data set has considered as the decision attribute. The decision table used for getting the reduct of attributes using RST has shown in table 3.

Step 3: Generation of Reduced set of Relevant Attributes through RST

RST can be used as a tool to discover data dependencies and to reduce the no. of attributes contained in the data set using the data alone, requiring no additional information. Rough sets theory defines three regions based on the equivalent classes induced by the attribute values: Lower approximation or Positive Region, Upper approximation or Negative Region and Boundary Region. Lower approximation contains all the objects, which are classified surely based on the data collected and Upper approximation contains all the objects, which can be classified probably, while the Boundary is the difference between the upper approximation and the lower approximation. Rough set reduct can be found out by using degree of dependency approach or using discernibility matrix method. Although it is guaranteed to discover all minimal subsets using discernibility matrix method, it is a costly operation. Again simplifying discernibility function for reduct is a NP-hard problem. Hence here it is preferred to use the dependency based approach. Using Rough Set Theory the reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same predictive capability of the decision feature as the original. Here we have first calculated the degree of dependency of each attribute and then the best candidate has chosen. This process has continued till the dependency of the reduct equals the consistency of the data set. The degree of Dependency can be found out by:

Υ = Sum of all Positive Regions / Total no of objects

So, from the Decision Table, the degree of dependency obtained using all the attributes is $\Upsilon_{A,B,C,D}(E) = (4+3+3)/10 = 10/10$

Degree of dependency of each attribute obtained is as follows:

$\Upsilon_{A}(E) = (0+0+3)/10 = 3/10$

$\Upsilon_{B}(E) = (0+0+3)/10 = 3/10$

$\Upsilon_{C}(E) = (0+0+0)/10 = 0/10$

$\Upsilon_{D}(E) = (3+0+4)/10 = 7/10$

As, the degree of Dependency of attribute 'D' is greater than other attributes, so Attribute 'D' is taken into consideration. Again the degree of dependency obtained, by taking the combination of attribute D and the other three attributes is as follows:

$\Upsilon_{A,D}(E) = (4+1+3)/10 = 8/10$

$\Upsilon_{B,D}(E) = (4+3+3)/10 = 10/10$

$\Upsilon_{C,D}(E) = (3+4+3)/10 = 10/10$

As the Degree of Dependency of {B, D} or {C, D} is higher than other and equal to the degree of dependency obtained with the original data set with four attributes, so both the attribute {B, D} and {C, D} can be taken as the reduct. So Minimal Reduced Attribute is R_{min} = {B, D} or {C, D}. The data set containing reduced set of relevant attributes has shown in table 4.

Step 4: Finding Max no. of possible clusters

There are many methods to find out maximum number of possible clusters for the given dataset. In this paper, the Rule of Thumb has applied to obtain the maximum no. of possible

cluster indexes. According to Thumb Rule $K \approx \sqrt{n/2}$ .Where, K = No. of Clusters and n = No. of data objects. Applying the Thumb Rule to the synthetic data set, the maximum no. of possible cluster indexes obtained is $K \approx \sqrt{10/2} \approx 2.0$.

Step 5: Finding the cluster index of each data object using the steps of genetic algorithms.

The basic steps of genetic algorithm for data clustering include encoding, fitness computation, selection, crossover and mutation. Each individual represents one feature subspace. Its fitness represents the clustering result with respect to the feature space that the individual represents. Larger the fitness, denser the data in such feature subspace and better the clustering results. The details are described below.

Individual Representation and Population Initialization

Individual representation or Encoding transforms one possible solution from solution space to search space which can be handled by GAs. The individuals are vectors of the solution space in the form of strings. One individual represents one possible solution to the problem. GAs can find the optimal solution or approximate optimal solution of the problem after applying a certain number of genetic operators on those individuals. There are two commonly used encoding methods: binary encoding and floating point encoding. Comparing with floating point encoding, the searching space of binary encoding is larger, moreover, the crossover and mutation implemented on it is more convenient. Therefore, binary encoding has adopted in this paper. Each attribute value has represented by an 8 bit string. Hence the data corresponding to each data object in the table 4 has represented by a 16 bit string.

Then the initial population has set up at random. At first, the K cluster centers encoded in each chromosome are initialized to K different randomly chosen features from the original feature space. Then, this process is repeated for each chromosome in the population. For the synthetic data set, initially the data object 4 (45, 20) and 8 (60, 35) has taken randomly as the cluster center 1 and cluster center 2 as K = 2.

Fitness Computation

The fitness computation process consists of two phases. In the first phase, the clusters are formed according to the centers encoded in the chromosome under consideration. After the clustering is done, the cluster centers encoded in the chromosome are replaced by the mean points of the respective clusters. The Fitness Function value calculated for the data objects of the synthetic data set has shown in table 5.

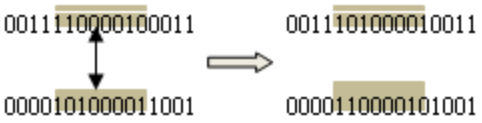
Selection

Selection process is used to get the optimum solution by preferring individuals with high fitness. By applying Tournament selection method in MATLAB, two random numbers between 1 to 10 has generated. Then, the no. having highest fitness function value is selected. By applying this procedure for 10 times the data objects selected has shown in table 6.

Crossover

Crossover is a probabilistic process that exchanges information between two parent chromosomes for generating two child chromosomes. In this paper two point crossover has applied to generate new offspring.

Example of two point crossover:



Mutation

Mutation process normally changes the structure of chromosomes, by negating a randomly chosen bit. As here binary representations of chromosomes are used, so a bit position (or gene) is mutated by simply flipping its value.

Example of Mutation:

The output of crossover and mutation applied to the synthetic data set has shown in table 7.

Termination Condition

In this paper the processes of fitness computation, selection, crossover, and mutation has executed for a maximum number of iterations, till the no of cluster centroids obtained is 2. The best string seen up to the last generation provides the solution to the clustering problem. Thus on termination, the new values generated provides the centers of the final clusters. The output of clustering with the reduced dataset has shown in table 8.

Lastly the comparative analysis of GA based data clustering with original attributes and reduced attributes has shown in table 9. Thus using Rough Genetic approach of data clustering, the cluster index obtained for the data objects is approximately same with cluster index obtained for the data objects using only the genetic algorithms with original attributes.

Conclusion

GA is the randomized search and optimization technique guided by the principles of evolution and natural genetics, and having a large amount of implicit parallelism. GA provides near optimal solutions for objective or fitness function of an optimization problem. This paper presents an efficient GA-based unsupervised clustering method based on the use of a binary chromosome representation and more effective versions of GA operators. As with the increase of the input sizes or dimensionality, the Genetic Algorithms may get computationally expensive, hence here we initially proposed to use an efficient method of Rough set theory for attribute reduction. The entire model has implemented on a synthetic data set. Using the degree of dependency based approach of RST, a reduced set of relevant attributes have generated, over which by applying the steps of genetic algorithm, the cluster index of each data object has obtained. Comparing the clustering result of the reduced data set with the original data set, it has been found approximately same. Hence the Rough Genetic Approach of data clustering may give better clustering accuracy by removing the irrelevant attributes through the steps of RST with reduced computational complexity.

References

[1] Yan Huang and Shulin Chen, “An Algorithm of Attribute Reduction based on Rough Sets”, International Conference on Computer Science and Software Engineering, Volume: 1, PP: 638-640, 2010.

[2] K. Thangavel, Qiang Shen and A. Pethalakshmi, “Application of Clustering for Feature Selection based on Rough Set theory Approach”, AIML Journal, Volume : 6(1), PP: 19-27,2006.

[3] Ujjwal Maulik, Sanghamitra Bandyopadhyay, “Genetic algorithm-based clustering technique”, Pattern Recognition, Volume: 33, PP: 1455-1465, 2000.

[4] Hwei-Jen Lin, Fu-Wen Yang and Yang-Ta Kao, “An Efficient GA-based Clustering Technique” , Tamkang Journal of science and Engineering, Volume :8(2), PP: 113-122, 2005.

[5] Rahul Kala, Anupam Shukla and Ritu Tiwary, "A Novel Approach to Clustering using Genetic Algorithm", International Journal of Engineering Research and Industrial Applications(IJERIA), Volume: 3(1), PP: 81-88, 2010.
[6] Wei Jian-Xiang, Liu Huai , Sun Yue-hong & Su Xin-Ning, "Application of Genetic Algorithm in Document Clustering" , Information Technology and Computer Science, Volume: 1, PP: 145 -148, 2009.
[7] Hao-jun Sun, Lang-huan Xiong, "Genetic Algorithm based High-Dimensional Data Clustering Technique", International Conference on Fuzzy Systems and Knowledge Discovery, Volume: 1, PP: 485-489, 2009.
[8] K. Thangavel and A. Pethalakshmi , "Feature Selection for Medical Database Using Rough System", AIML Journal, Volume: 6(1), PP: 11-17, 2009.
[9] M. Zhang and J. T. Yao, "A Rough Set based Approach to Feature Selection", North American Fuzzy Information Processing Society (NAFIPS), Volume: 1, PP: 434- 439, 2008.
[10] Liu Peng, Wang Qing, Gu Yujia, "Study on Comparison of Discretization Methods", International Conference on Artificial Intelligence and Computational Intelligence, Volume: 4(1), PP: 380-384, 2009.
[11] Kotsiantis Sotiris and Kanellopoulos Dimitris, "Discretization Techniques: A recent survey", GESTS International Transactions on Computer Science and Engineering, Volume: 32(1), PP: 47-58, 2008.
[12] Sellappan Palaniappan, Tan Kim Hong, "Discretization of Continuous Valued dimensions in OLAP Data Cubes", International Journal of Computer Science and Network Security, Volume: 8(11), PP: 116-126, 2009.

Table 1 Synthetic Dataset

Object No.	A	B	C	D
1	10	15	30	40
2	40	20	40	50
3	15	60	35	15
4	20	40	45	20
5	50	30	20	60
6	30	35	15	30
7	45	50	50	45
8	35	25	60	35
9	25	10	10	25
10	60	45	25	10

Table 2 Output of Discretization

Object No.	A	B	C	D
1	2	2	1	3
2	1	2	1	1
3	2	3	1	2
4	2	1	3	2
5	3	1	2	1
6	1	1	2	3
7	3	3	3	1
8	1	1	3	3
9	1	2	2	3
10	3	3	2	2

Table 3 Decision Table Corresponding to the Synthetic Dataset

Object No.	A	B	C	D	Dec. Attr.(E)
1	2	2	1	3	1
2	1	2	1	1	2
3	2	3	1	2	2
4	2	1	3	2	1
5	3	1	2	1	2
6	1	1	2	3	1
7	3	3	3	1	2
8	1	1	3	3	1
9	1	2	2	3	1
10	3	3	2	2	2

Table 4 Data Set with Reduced Set of Relevant Attributes

Object No.	C	D
1	30	40
2	40	50
3	35	15
4	45	20
5	20	60
6	15	30
7	50	45
8	60	35
9	10	25
10	25	10

Table 5 Binary Representation and Fitness Computation

Object No.	Binary Representation in 16 bits	Fitness Function Value
1	0001111000101000	(0.023, 0.040)
2	0010100000110010	(0.040, 0.023)
3	0010001100001111	(0.028, 0.040)
4	0010110100010100	(0.067, 0.067)
5	0001010000111100	(0.015, 0.015)
6	0000111100011110	(0.013, 0.067)
7	0011001000101101	(0.067, 0.028)
8	0011110000100011	(0.067,0.067)
9	0000101000011001	(0.011,0.067)
10	0001100100001010	(0.018, 0.028)

Table 6 Output of Selection Procedure

Random no. s generated between 1 to 10	Data object with highest fitness value	Corresponding Binary representation
8, 9	8	0011110000100011
1,9	9	0000101000011001
6,1	6	0000111100011110
3,5	3	0010001100001111
9,10	9	0000101000011001
1,10	1	0001111000101000
9,5	9	0000101000011001
8,1	8	0011110000100011
4,9	4	0010110100010100
7,9	7	0011001000101101

Table 7 New Data Values Obtained by applying all the steps of Genetic Algorithm once

Old value	O/P of Crossover	O/P of Mutation	New Value
30,40	0011101000010011	0011101000010010	(58,18)
40,50	0000110000101001	0000110000101000	(12,40)
35,15	0000001100001110	0000001100001111	(3,15)
45,20	0010111100011111	0010111100011110	(47,30)
20,60	0000111000101001	0000111000101000	(14,40)
15,30	0001101000011000	0001101000011001	(26,25)
50,45	0000110000101001	0000110000101000	(12,40)
60,35	0011101000010011	0011101000010010	(58,18)
10,25	0010001000100100	0010001000100101	(34,37)
25,10	0011110100011101	0011110100011110	(61,28)

Table 8 O/P of GA based Data Clustering with Reduced Attributes

Object No.	Dataset (C,D)	New Value	Cluster Index
1	(30,40)	(58,19)	1
2	(40,50)	(58,19)	1
3	(35,15)	(58,19)	1
4	(45,20)	(42,20)	2
5	(20,60)	(58,19)	1
6	(15,30)	(42,20)	2
7	(50,45)	(58,19)	1
8	(60,35)	(58,19)	1
9	(10,25)	(42,20)	2
10	(25,10)	(58,19)	1

Table 9 Comparative Analysis of GA based Data Clustering with Original and Reduced Attribute Sets

Object No.	Cluster index of each object with Original Attributes	Cluster index of each object with Reduced Attributes
1	1	1
2	1	1
3	1	1
4	2	2
5	2	1
6	2	2
7	2	1
8	1	1
9	2	2
10	2	1