International Journal of Engineering & Technology, 7 (4.30) (2018) 550-552

International Journal of Engineering & Technology

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

Research paper

A cluster Analysis for Binary Data Using Genetic Algorithms

Sabariah Saharan[1], Wong Yu Xian[2], Roberto Baragona[3*]

[1,2]Department of Mathematics and Statistics,
Faculty of Applied Science and Technology,
Universiti Tun Hussein Onn Malaysia, Pagoh Educational Hub,
84600 Pagoh, Johor, Malaysia
[3]Department of Communication and Social Research,
Sapienza University of Rome, Italy
**Corresponding author E-mail: roberto.baragona@uniroma1.it*

Abstract

This research was initially driven by the lack of clustering algorithms that focus on binary data. A promising technique to analyze this type of data, namely Genetic Clustering for Unknown K (GCUK) became the main subject in this research. GCUK was applied to cluster four binary data and there is a presence of an imbalanced data in one of the data sets. The results show that GCUK is an efficient and effective clustering algorithm compared to K-means. The other contribution is the capability of GCUK for clustering the unbalanced data. Standard clustering algorithms cannot simply be applied to this type of data sets as it can cause a misclassification results.

Keywords: *Binary Data; Clustering; Genetic Algorithms.*

1. Introduction

The big volume of data that are collected by an firms or individual has urge the researchers to explore more techniques to analyze this data. Clustering is one of the most popular machine learning technique that can be used to process this big data set. Clustering is difference from classification where it finds natural group of objects based on the similarities or relationship. The similarities are measured by the distance measurement and the closest distance will be put in the same group. While for the classification, it is a learning method for predicting the class object from pre-classified label. Usually the validation of clustering is difficult because the number of cluster K is not known prior. So, it can be a challenging problem for a researcher to find a suitable K for the data set [1]. The most common algorithms being used by most researchers to cluster data is K-means algorithm [2]. This well-known partitioning algorithm uses an iterative process to cluster the data. However, it is using a single point as their searching space, which makes it easily to stuck in local optima [3].

To overcome this problem a promising technique namely Genetic Algorithms (GAs) was used in this study. This algorithm has been used for many applications such as optimization, engineering, biological sciences and clustering. Compared to traditional algorithm, this technique using a population points as a search space, which can avoid to be trapped in a local optimum. GAs also have advantage for clustering task, where it can solve the number of clusters and allocate these items to its cluster solve simultaneously [4].

One of the efficient and effective GAs based clustering is Genetic Clustering for Unknown K (GCUK). It was originally proposed by [5] and produced the best results to cluster the large numerical data sets compared to the other clustering algorithms. In this study, GCUK algorithm with some improvement proposed by [6] was applied to cluster the binary data sets.

Binary data has a special place in the system data analysis. This type of data usually represented as a binary vector to indicates whether a given term or point was present or not. Market basket transaction and document clustering is some of the example of binary applications. In this data sets, the data points were represented by the variables and are much less than the point coordinates. Hamming distance was used as an acceptable measure on binary points.

2. Methodology

Among GA based clustering algorithm, GCUK is one of the most efficient method that used single objective optimization [7]. GCUK was proposed by [5] and [8] to test the effectiveness of GAs. In GCUK, the chromosome that is representing any possible solution was coded as a string of the maximum number of clusters, K. These characters can be either coordinates of the data points or a symbol of # (do not care) which represent the unassigned genes. Let $k_{\min} = 2$ and $k_{\max} = 8$ with the chosen number of cluster is $k = 3$. The chromosome i may look like following:

$$(13.1,10.2),\#,\#,(14.1,2.9),\#,\#,(14,5.6),\#$$

However, because of the real number was used as a string representation to encode the centers, the cost of computation time for floating-point computing is high. It means that the cluster center should be recomputed every time to check for the fitness which requires a high computational time.

Thus, in this paper, the string represented by the cluster center and not the real number as the previous research. In GCUK, a chromosome is representing the cluster center. Then the coordinate of the clusters is represented by the values of the factors and it can be considered as a length of the dimension of the data set d .

Let say that the number of clusters is k , then the length of the chromosome is equal to kd . The differences between the cluster centers and the data points was calculated by using the Hamming distance. More formally, the distance between two strings A and B is $\sum|A_i - B_i|$. Unlike the original GCUK in [6] where the number of clusters were automatically given by the algorithm, here in this study, it was pre-specified by using Silhouette index. This is to make sure that the best number of clusters have been chosen to do the analysis and to avoid having any empty cluster.

Population Initialization: Assume that the population size s **,** and the initialization i^{th} chromosomes, Ch_i **,** in the population of $(i = 1, 2, \cdots, s)$ and k_i is fixed by $k_{\min} = k_{\max}$. For the binary problem, the gene is corresponding to the index of each of the cluster centers. This was represented as “1”, while the remaining gene was represented as “0”. For example, if the length of the data is $l = 8$ and four cluster centers were chosen randomly from the data set which have indices of 3,6, and 8 respectively, then the chromosome Ch_i is 00100101. In this study, there is no genes are left unassigned (#) as in the original GCUK by [6].

Fitness Function: Only one fitness function was used to be met in here. In this study, GCUK uses the intra-cluster variances as the fitness function where the lowest values will indicate the better results. The sufficient statistics $N_j = \sum_{t_i \in C_j} 1$, is the number of cases in cluster $j, j = 1, \ldots, k$, and M_j, is the vector whose entries are the sums of 1 for each factor of the data points in cluster C_j are needed. For each cluster the variance of the proportion $C_j = M_j / N_j$ is equal to $R_j = C_j(1 - C_j)$. The intra-cluster variances can be written as follows:

$$q(R,W) = \sum_{j=1}^{k} W_j \sum_{i=1}^{d} R_{ij} \tag{1}$$

Given that $W_j = \frac{N_j}{n}$. The lower the $q(RW)$ value indicated the better the clustering results in achieving objective function.

The following operators are the essential process in GAs as implementation of GCUK to deal with binary data sets. In this algorithm, roulette wheel selection (RWS), single point crossover and uniform mutation was used as a genetic operator.

Selection: The conventional proportion (RWS) was used. This selection method is a circular wheel where the regions in the wheel is represent the fitness values for the chromosomes. The probability of each of the chromosome is a proportion to its fitness. The largest region, represent the fittest chromosome and have greater chances to be selected, compared to the smaller region (weakest chromosome). The wheel will be spun N times and the pointer will stop at any region. These selected chromosomes will then replace the whole population, P and become a new population P' .

Crossover: The selected chromosome then will be paired at random and each pair has a fixed probability p_c , to generate another two children. A single point crossover was used to exchange part of two chromosomes (parents) to yield better children. This process is to explore further regions of the solution space and to obtain a better chromosome.

$$\text{Parent} \quad \begin{matrix} 1011 \,\|\, \mathbf{0100} \\ 1100 \,\|\, \mathbf{1101} \end{matrix} \quad \rightarrow \quad \text{Children} \quad \begin{matrix} 1011 \,\|\, \mathbf{1101} \\ 1100 \,\|\, \mathbf{0100} \end{matrix}$$

For example, two chromosomes from parents, will exchange all the alleles at their 5th genes. Then, all the alleles after 4th change each other and produce two new children.

Mutation: Mutation process is a process where a small random tweak in a chromosome. In this study, each of the chromosome will be mutates with probability p_m . Assume that v is a coordinate. Then, when a mutation occurs, v will becomes either $v = \pm 2\delta$ if $v = 0$ or $v(1 \pm 2\delta)$. δ is a number generated from the uniform distribution [0,1] and the sign + and – will occurs with an equal probability. However, if the mutation result has exceeded the extremes interval of [0,1], it will reset either to 0 if less than zero or to 1 if greater than one. In this study, the uniform random mutation was used and each gene in the chromosome will be flipped with a pre-specified probability p_m . Then, it will be replaced the previous population P' with a new population P'' .

Termination process: There is no fully conclusive until now on when to stop the process of GCUK. Some of the idea to terminate the process is, to stop when the upper limit of the generations has been reached [9]. Thus, in this study, when the process is achieved the maximum number of iterations, then the process will be executed by assuming that it will give the smallest fitness value.

3. Analysis and results

Three real life data sets were retrieved from UCI Machine Learning Repository, which were car evaluation, bank marketing and nursery application data. The car evaluation data set contains information about the performance of the car. It has 1728 observations with 21 criteria to be clustered. For bank marketing, this data has 45211 observations and it gave an information about the marketing campaign done by the Portuguese Banking institution. The nursery data had observations of 12960 with eight groups of categorical variables. It is explaining about the criteria needed for choosing the children before they can enter the primary school. While for road traffic accidents data, was taken from the government website of UK (Department of Transportation, 2016). The road traffic accidents data had an imbalanced class problem where the class of positive and negative are not evenly distributed. In this case, the positive class (fatality case) become a minority class while the negative (non-injured case) is a majority. Standard clustering cannot be simply applied to this kind of data set, as it will cause misclassification results.

Table 1 shows the average values of Silhouette index for all the data sets to find the best number of clusters, K .

Table 1: Silhouette index for different data sets

Car	Bank	Nursery	Road Accidents
$k = 3, s_i = 0.676$	$k = 5, s_i = 0.143$	$k = 7, s_i = 0.059$	$k = 5, s_i = 0.823$
$k = 4, s_i = 0.719$	$k = 6, s_i = 0.491$	$k = 8, s_i = 0.104$	$k = 6, s_i = 0.838$
$k = 5, s_i = 0.310$	$k = 7, s_i = 0.191$	$k = 9, s_i = 0.089$	$k = 7, s_i = 0.751$

From Table 1, the suggestion K for car data is when $k = 4$ with the highest value of Silhouette index is $s_i = 0.719$. The best choice of K for bank data set is $k=6$ with $s_i = 0.491$. While for nursery data, Silhouette index gave a suggestion that the number of clusters K is $k=8$ and the average value is $s_i = 0.104$. Lastly, the suitable number of clusters for road traffic accidents is when $k=6$ which gave the Silhouette index, $s_i = 0.838$.

After the process of selecting the best number of clusters, the next process is to set the parameter setting for each of the data set. GAs was built with different number of parameter settings to find a better solution. The following parameters were used in this study; number of population size (*NIS*), maximum number of generation (*MG*), crossover probability (p_c) and mutation probability (p_m). The algorithm need to run several times until the best solution was performed. The results of the different parameters setting are shows in Table 2 below.

Table 2: $q(RW)$ for different data sets with different parameters

Data sets	Parameters				$q(RW)$
	NIS	*MG*	p_c	p_m	
	80	100	0.90	0.001	200.00
Car	100	150	0.80	0.01	196.57
	150	180	0.80	0.001	208.89
	30	150	0.70	0.001	524.77
Bank	50	80	0.80	0.01	539.91
	80	150	0.80	0.001	565.378
	40	90	0.80	0.001	2266.10
Nursery	30	90	0.70	0.001	2301.20
	40	90	0.80	0.01	2350.10
	25	90	0.90	0.05	1308.70
Accidents	30	100	0.80	0.01	1280.80
	50	80	0.80	0.01	1350.20

After several round of testing with different parameters, the best parameters for car data is when *NIS* =100, *MG* = 150, p_c =0.80, p_m = 0.001 and $q(RW) = 196.57$ which is the lowest value among the others. For bank data set, the suggestion parameters were *NIS* =30, *MG* = 150, p_c =0.70 and p_m = 0.01 and the value of $q(RW)$ is 524.77. While for nursery, the selection of the setting was when *NIS* =40, *MG* = 90, p_c =0.80 and p_m = 0.001 with the value of $q(RW)$ is equal to 2266.10. As mentioned before, the road traffic accidents data set is a special one due to the presence of imbalanced classes. For this data set, the parameters of *NIS* =30, *MG* = 100, p_c =0.80 and p_m = 0.01 was chosen as it gave the lowest value fitness value, $(q(RW) = 1280.80)$. By assuming that all these fitness value were enough to compete with K-means, the process of searching the best parameters setting was stopped. It also due to the power of computing engine to do the iterations and limitation of time.

The standard clustering, K-means algorithm has been ran to compared and checking the validity of the results provided by the GCUK. To check the performance of this algorithm, K-means was run with the same number of clusters, K and compared the fitness value. The results from the GCUK and K-means algorithm are reported in Table 3 below.

Table 3: Comparison of $q(RW)$ between GCUK and K-means algorithm

Data Sets	Intra-cluster variance, $q(RW)$	
	GCUK	K-means
Car Evaluation	196.5714	225.7143
Bank Marketing	524.7740	553.7452
Nursery	2266.000	2446.7000
Road traffic accidents	1280.8000	1390.6000

From Table 3, it shows that the value of $q(RW)$ for all the data sets using GCUK is lower than K-means algorithms. It proved that GCUK is competent to be efficient to cluster the binary data sets. It also showed that this algorithm gave a good performance when clustered the imbalanced class data set.

4. Conclusion

In this study, GCUK showed a satisfactory capability to handle large binary data sets. A comparison with K-means algorithm seems to validate these results. There maybe at least one issue that seems relevant as further research topics. The concerned is with the data streaming such as road traffic accidents, the stability of the results when a new data set updated in the system may affect the present conclusions [10]. However, for road traffic accidents, more detailed information and involving the time series data will give a researcher to investigate further on this issue in regard to traditional clustering algorithms.

Acknowledgement

Our thanks to Universiti Tun Hussein Onn Malaysia for providing us a financial support to finish this research. This study was supported by TIER 1 Project Grant Vot. U900.

References

[1] Hruschka ER, Campello R, Freitas AA & de Carvalho A (2009), A Survey of Evolutionary Algorithms for Clustering/ Systems, Man, and Cybernetics, Part C: Applications and Reviews. *IEEE Transactions* 39(2), 133-155.

[2] Jain AK (2010), Data clustering: 50 years beyond K-means, *Pattern Recognition Letters* 31(8), 651-666.

[3] Ordonez C (2003), Clustering binary data streams with K-means. In DMKD03: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 12-19

[4] Baragona R, Battaglia F, Polu, *I. Evolutionary Statistical Procedures*, Springer, Berlin and Heidelberg, (2011).

[5] Bandyopadhyay S, Maulik U (2002), Genetic Clustering for Automatic Evolution of Clusters and Application to Image Recognition. Pattern Recognition, 35, 1197-1208.

[6] Saharan S & Baragona R (2013), A New Genetic Algorithm for Clustering Binary Data with Application to Traffic Accidents in Christchurch. *Far East Journal of Theoretical Statistics* 45(1), 67-89.

[7] Lin HJ, Yang FW, Kao YT (2005), An Efficient GA-based Clustering Technique. *Tamkang Journal of Science and Engineering* 8(2), 113-122

[8] Maulik U, Bandyopadhyay S (2000), Genetic Algorithm-based Clustering Technique. *Pattern Recognition* 33(9), 1455-1465.

[9] Safe M, Carballido J, Ponzoni I & Brignole N (2004), On Stopping Criteria for Genetic Algorithms. *Advances in Artificial Intelligence*, 405-413.

[10] Milligan G, Cheng R (1996), Measuring the influence of individual data points in a cluster analysis. *Journal of Classification* 13(2), 315-335.