A NEW METHOD TO FEATURE SELECTION IN HIGHDIMANTIONAL DATA SETS FOR DATA MINING

S. Alavi, PhD
Associate Professorof the Shahid Chamran University of Ahvaz-Ahvaz

H. Ahmadi
Research Scientist of the Shahid Chamran University of Ahvaz-Ahvaz

Shahid Chamran University of Ahvaz-Ahvaz, Iran
E-mail: se.alavi@scu.ac.ir, haloojhen@gmail.com

Abstract. A method of feature selection by using clustering graph is presented. in the suggested method, using a detection algorithm communities at first, the basic characteristics divided into some clusters. Then, by applying genetic algorithms and using a KNN (K nearest neighbor) classification algorithm a feature selection method based on covering solution is presented. The performance of the suggested method compared with the most recognized and the most recent feature selection methods applied on different classificators. The results showed that the suggested method both in terms of time and classification accuracy has a proper function.

Keywords: feature selection, covering solution, genetic algorithms, data mining, graph clustering, community detection

Introduction. In accordance with the rapid growth of technology, databases and computers; data are accumulated faster than the processing capacity of human beings. These high-dimensional data collections greatly reduce data mining tools' performance [1]. High-dimensional data collections reduce the performance of the classifier of two sides. On the one hand, by increase in the size of the data, volume calculations increase and on the other hand, the model built based on high-dimensional data has low level of interoperability [2]. Two major strategies to reduce the size of the data collections are provided; Feature Extraction and Feature Selection. Feature selection is one of the most important steps in the pre-processing step of data mining process. In the feature electoral process, by eliminating unrelated features and redundancy; a subset of the original features selected. In this study, the combination of communities detection algorithm with genetic algorithm is used for feature selection. In this method for clustering features, a graph-based clustering algorithm for clustering is used. This feature selection method, for evaluating candidate subset in the process of feature selection, uses KNN classification algorithm.

Past works. Based on the evaluation criteria, feature selection methods generally divided into four solution; covering, filtering, hybridding and embedding. In [3] for

feature selection in matters of medical diagnosis, a combination of particle swarm algorithm and rough set theory is used. In many methods of feature selection, genetic algorithm was used to for finding optimal subset. Stefano and his colleagues [4] to improve the classification performance in handwriting recognition have offered an effective feature selection method. In genetic algorithms provided by them, to assess the candidate subsets, a separability index is used. Kabir and his colleagues [5] have proposed a hybrid genetic algorithm for feature selection. The most important aspect of their presented algorithms is the automatic designation of subset size and the selection of a subset features with the smallest size. Zhao and colleagues [6] on the idea of using clustering features offered an unsupervised feature selection method. One of the active and challenging research areas in social network discussions is the subject of recognizing communities. The main purpose of recognizing communities is that the similar nodes set in a cluster and the relation between nodes within each cluster become denser than the nodes of different clusters. One of the fast and efficient algorithms for detecting communities is *Louvain* algorithm that carries out graph clustering using the maximum modular function [7].

The suggested method. The proposed method consists of four stages; reducing the size of data collection, graphic representation of the problem, feature clustering and searching optimal sub-set based on genetic algorithm. In the first stage of suggested method, if the numbers of original features are more than, suitable features should select and the rest must be deleted. In order to clustering the features with the communities' detection algorithm, the problem represents in a complete weighted, undirected graph. In this paper, to calculate the similarity between features, the absolute value of the correlation coefficient and for clustering features of a community detection algorithm called Louvain is used (figures 1).

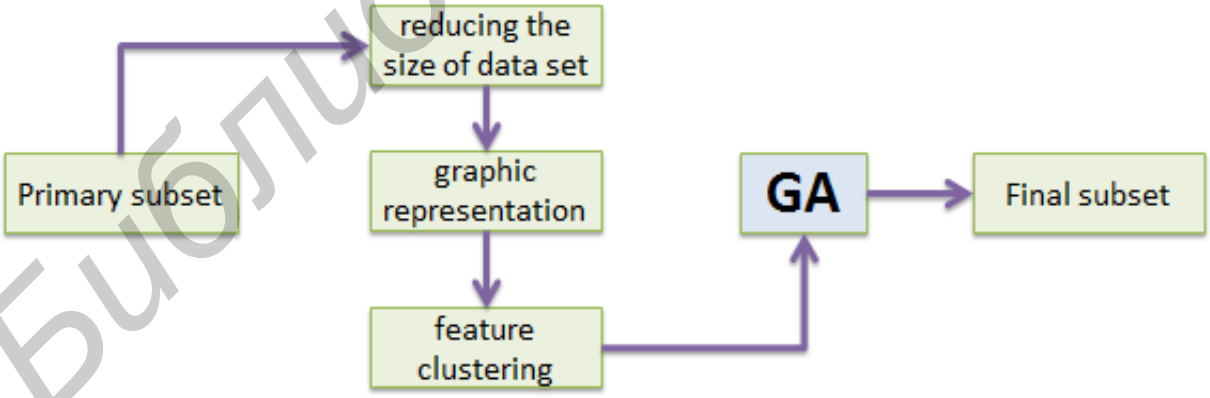

Fig. 1. Suggested method general Scheme

In order to increase the efficiency of the suggested method, before running communities detection algorithm, by applying a threshold on the graph edges, those edges that weighs less than the threshold θ are deleted. In the end, by using a genetic algorithm and with the help of clustering features, optimized features subset is searched.

For this purpose, a restoration operator at the end of each iteration of genetic algorithm applied to amend the chromosomes, in a way that from each cluster, at least ω features selected. The fitness function in genetic algorithm is calculated using relation (1).

$$J\left(FS^k(t)\right) = \frac{CA\left(FS^k(t)\right)}{\frac{2}{|FS^k(t)| * (|FS^k(t)| - 1)} \sum_{F_i, F_j \in FS^k} Sim(F_i, F_j)} \quad (1)$$

here $CA\left(FS^k(t)\right)$ is classification accuracy for chosen feature subset $FS^k(t)$ on KNN classifier,

$FS^k(t)$ chosen feature subset size and (F_i, Fj) Likeness among feature F_i, F_j .(k=3)

The most important step in suggested method is the restoration in genetic algorithm that is applied on all chromosomes. This operator is done in a way that ω number of features is surly selected from each cluster. To select and delete features from a cluster two different strategies of random restoration and preferential restoration exists, that is used on the preferential restoration of selection possibility or deleting features in the process of restoration according to Fisher rating.

Results. In this thesis, the Colon data set with 2000 features and 2 classes is used. Table 1 shows the values of various parameters of the suggested method. Methods compared in this part are compared with each other based on three criteria of; the number of selected features, classification accuracy and execution time. It is worth mentioning the suggested method is displayed by GA (genetic algorithm).

Table 1. Parameters

Parameter	**Notation**	**Value**
Number of iteration	A	50
Population size	P	Number of features
Size of reduced features	N	100
Crossover rate	Crs	0,8
Mutation rate	Mut	$1/A$
Threshold for remove edges	θ	0,5
Number of selected feature from each cluster	ω	2

Table 2. Average size of the subset selected by the suggested method in comparison with the other methods

Dataset	Feature Selection Method				
	GA*	**HGA [5]**	**ACO [8]**	**PSO [9]**	**All features**
Colon	10,6	12,8	11,9	96,9	2000

Table 3. The mean and variance of classification accuracy in the suggested method in comparison with other methods of feature selection on SVM classifier. (Here Acc shows classification accuracy and Std shows standard deviation of ten independent performances.)

Dataset		Feature Selection Method					
		GA* Random	**GA* *Score***	**HGA [5]**	**ACO [8]**	**PSO [9]**	**All features**
Colon	Acc (%)	1,51	85,78	82,58	83,79	48,88	80,22
	Std	84,73	1,79	1,88	1,80	1,68	1,28

Table 4. Runtime in minutes in the suggested method in comparison with other methods

Dataset	Feature Selection Method				
	GA* Random	**GA* *Score***	**HGA [5]**	**ACO [8]**	**PSO [9]**
Colon	0,91	0,96	0,93	0,81	374,82

The results of this table shows the suggested method with random restoration and the suggested method with preferential restoration have a running time shorter than other methods of feature selection.

Conclusion. The assessment of the proposed method and its performance comparison with other methods shows that the proposed method has appropriate performance and in the most data sets has the best performance among different ways. Moreover, the results of comparing different strategies of random restoration and preferential restoration showed that in the majority of data collection, the suggested method of preferential restoration has a better performance.

Future works. In this paper a framework for graph representations was presented. So we can make use of researches done in done in the other areas of science, such as social networks and graph theory and Identified more accurately the relationship between the features. For instance, the techniques such as dissemination of information, popularity and dominant-Set that widely applied on social network analysis can be used.

Reference

[1]. Liu. H. “Feature Selection; An Ever Evolving Frontier in Data Mining” . Computer Science and Engineering, Arizona State University, USA, 2002.

[2]. Cadenas, J.M., M.C. Garrido, and R. Martínez,“Feature subset selection Filter–Wrapper based on low quality data”. Expert Systems with Applications, 2013. 40 (16): p. 6241-6252.

[3]. Inbarani, H.H., A.T. Azar, and G. Jothi, “Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis”. Comput Methods Programs Biomed, 2014. 113(1): p. 175-85

[4]. C. De Stefano, et al., “A GA-based feature selection approach with an application to handwritten character recognition”. Pattern Recognition Letters, 2014. 35: p. 130-141.

[5]. Md. MonirulKabir , Md. Shahjahan, and K. Murase., “A new local search based hybrid genetic algorithm for feature selection”. Neurocomputing, 2011. 74(17): p. 2914–2928

[6]. Xi Zhao, W.D., Yong Shi, “Feature Selection with graphival modle by Information Gain”. Procedia Computer Science, 2012. 18: p. 43-64.

[7]. Vincent, B, Jean.G., Renaud. L, and Etinne. L, “Fast unfolding of communities in large networks”. Journal of Statistical Mechanics: Theory and Experiment, 2008. 10008: p. pp. 1–12.

[8]. Kabir, M.M., M. Shahjahan, and K. Murase, “A new hybrid ant colony optimization algorithm for feature selection”. Expert Systems with Applications, 2012. 39(3): p. 3747-3763.

[9]. S. Theodoridis and C. Koutroumbas, “Pattern Recognition”, 4th Edn. Elsevier Inc, 2009.