Presentation of Hybrid Genetic Algorithm for Effective Feature Selection

Hamideh Ganji-Arjenaki, Mohammad-Hossein Nadimi-Shahraki, Nasim Nourafza

Abstract—Many sciences encounter large volume of information with advancement of data collection and storage capabilities in recent decades. Nowadays, substrate data creates new challenges in data analysis while traditional statistical methods are not responsible for these data analysis due to the increase in number of observations and variables. Feature selection was used for solving this problem. In selection feature, the best combination of features is surveyed and it requires time and high processing. This paper tried to use the combination of genetic algorithm and artificial neural network for solving this problem. A single-layer Perceptron is topology of artificial neural network. The results of the study showed that the best combination from set of features was found in the shortest time and more optimal way.

Index Terms— Feature selection, Genetic algorithm, Artificial neural networks.

1 Introduction

Researchers in different fields such as engineering, astronomy, biology, and economics face with more and more observations. Traditional statistical methods missed their efficiency because of two reasons. The fist reason is the increase in number of observations.The second reason which is of utmost importance, is the increase of variables in one observation. Therefore, new data analyses face with serious challenges [1]. For example, the varied experiments should be performed on the patients in order to diagnose disease in medical field. This work is in company with cost and side effects on the patients. At last, large volume of information is available that all of them are not used in diagnosing. There is a high probability that they cannot be used for disease diagnosis [2], [3], [4].

Effective feature selection not only reduces cost and time of disease diagnosis, but also achieves the best combination. The present study consists of the following sections: reduction of data dimensions, genetic algorithm, and advantages of using genetic algorithm in feature selection, results, and future work.

2 Feature Selection

Substrate data have many dimensions and create many computational challenges although they generate opportunities. One of the problems in data with numerous dimensions is that all features of data for finding knowledge hided in data are not vital. So, reduction of data dimensions is still one of important topic in many fields. Data dimension reduction methods are divided for two groups: methods based on feature extraction and methods based on feature selection [1], [5]. The aim of this paper is to present an effective method for feature selection.

Feature selection is finding a subset from features with the minimum numbers. So, it should include sufficient information. The main purpose of all algorithms and feature selection method is this subset.

Fakunaga and Narewenda (1977) presented a definition and they brought up selection of a subset with M elements from among N features, while M is smaller than N (M < N) and value of criterion function for subset should be more optimal than other subsets. Various feature selection methods try to find the best subset between 2 subsets of candidate. In all of these methods, the subset is selected as an answer that can optimize value of criterion function. Each method tries to select the best features, however, with the respect the breadth of possible answer and the increase of sets of answer exponentially with N, finding optimal answer is difficult and N medium and large is very costly [6].

Feature selection process includes four steps: production subset, evaluation subset, stopping criterion, and validation results [7].

2.1 Genetic Algorithm

John Holland from Michigan University proposed the use of genetic algorithm in optimizing engineering. The main point of this algorithm is transmission of inherited attributes by genes. Set of features are transferred in human beings to next population by their chromosomes. Each gene is representative of one attribute in these chromosomes.

Two events were happened for chromosomes simultaneously. The first event is mutation and another one is crossover [8], [9].

3 Related Work

As mentioned before, the experts try to reduce the data dimension in different fields and select effective features because of the increase in the number of features and dimen-

- *Faculty of Computer Engineering, Najafabad branch, Islamic Azad University, Najafabad , Iran .*
  E-mail:. h.ganji@sco.iaun.ac.ir
  E-mail: nadimi@iaun.ac.ir
  E-mail: n_noorafza@yahoo.com

sions. There are applied researches in medical field about this issue. Some of these studies are investigated on feature selection and some other investigated on feature extraction.

Arsalan and Turkoglu (2002) presented one system for diagnosing heart valve disease by using data diming and its techniques. In this system, the interpretation of Doppler signals of the heart was carried out on the basis of pattern recognition. Disease diagnosis was done based on the feature extraction from waveform and classification of features was done with neural network of propagation [10].

Damtew (2011) used feature selection for predicting of heart disease.In this method ,the preprocessing was done on the data and then, features were selected by best first search method. After that, they were classified by J48 method in order to determine the percentage of accuracy in predicting [11]. In this search method, all subsets were searched by feature selection method and the answer did not far from algorithm. On the one hand, in this method was checked all subsets. So, That can be time-consuming especially when the number of features are large.

Lee et al. (2008) used solution of feature extraction and classification for predicting heart disease. Greedy hill climbing was used for feature extraction and CPAR and SVM methods was used for classification [12].

R. Alizadeh et al. was done some researches on three vessels of LCX, LAD, and RCA for determining cardiovascular. The use of feature selection method and classification for evaluating selected features was the solution of this study. Also, the gain ratio was used for feature selection. The effective features were recognized based on its value. C4.5 Classifier was used based on obtained features [4].

H. Yan et al. searched about use of feature selection for diagnosing of heart. The aim of this paper was to select effective features. Algorithm genetic was used for diagnosing of heart disease [13].

4 The Proposed Method

First, a certain number of inputs (x1, x2, …, xn) from sample space of X are selected. Then, they are put in a vector x=(x1, x2, ….) that xn is called chromosome. The group of chromosomes is called population. Each chromosome grows in each period and develops based on specific rules of biological evolution.

There is a fitness function for each chromosome xi which is called f (Xi). The stronger elements or chromosomes value of optimization is competence of chromosomes. The stronger elements or chromosomes. They have more chance to survive in other periods and they can be reproduced.

Weak elements are ruined. In fact, genetic algorithm retains inputs that are near to more optimal answer and abandon others. Birthday is another important step in this algorithm and accrues in this period. The contents of two chromosomes are combined with together in process of generation in order to generate two new chromosomes. This rule allows two the best parents combine with together for generating better offspring. Furthermore, a series of chromosomes may be mutated in each period.

In this solution in order to create new generation, firstly, the new generation are formed by roulette wheel selection, crossover and mutation operation. In crossover operator use one point, two point and three points.

Fitness function is used for evaluating new generation which the amount of competency is determined in that chromosome. In this solution, the artificial neural networks with the topology single layer of Perceptron is used for fitness function. Therefore, the primary data set are filtered with respect to set of generated feature (chromosomes which are surveyed) and they are classified with artificial neural network. The higher percentage of accuracy shows the better set of features.

5 Discussion

Genetic algorithm does not require certain mathematic and it tries to solve optimization problems without paying attention to inner performance. This algorithm is able to solve any limitations (such as linear or nonlinear) which is defined on continuous, discrete or mixed search space.

Performance of this algorithm has been demonstrated experimentally. Structure of genetic algorithm operations makes this algorithm to act more successful in finding optimal answers. In traditional methods, the search is done with adjacent points by comparison and move to the points with relative optimization. Genetic algorithm has high flexibility in combination with innovative techniques. So, it can solve the problem effectively [9].

Genetic algorithm can solve different problems by coding in the form of chromosome.Structure of genetic algorithm presents tools for optimizing parameter solution in a specific problem. This is an easy and understandable method that math requirements are very low (or even no). This tool can be simulated easily by mediums [14].

This algorithm is considered an effective tool for the search if there is a little background knowledge about the problem and evaluating quality of selected samples is not available.

Genetic algorithm is a random search model and produces and evaluates set of various features in a short of time. In this algorithm, the number of subsets can be increased in several times with trifle changes. If this method is done with algorithms of other selected features, it requires time and high calculations.

Genetic algorithm is more common and useful method among other methods for finding suitable feature selection because of two reasons. The first reason is having high power in selecting varied features and the second one is having high speed. Use of genetic algorithm makes a rapid move in space of problem. This algorithm has high capability in problem solving [15]. If feature selection is based on fitness function, the subsets having negative effect are not selected as the final subset. Genetic algorithm has more power in finding answer of problem, especially when space of mode is big.

With selection of fitness function, the subset having higher accuracy of classification and lower cost, has a bigger value of fitness function. The final value of fitness function is considered as a ranking for each generated subset and the subset that acquires higher rank, is selected as the best subset.

Structure of problem and capability of genetic algorithm in suitable modeling is another reason in selection of genetic algorithm for problem solving and feature selection.

In addition to above-mentioned advantages for genetic algorithm, previous researches applied this technique for disease diagnosis and achieved good results. This is another reason for reliability in using this algorithm [3], [16].

If the answer is among chromosomes that are in lower rank with respect to fitness function, the chromosomes will have chance of selecting by the use roulette wheel operator [17], [18].

The proposed solution can be used in other problems easily and can be taken advantage of its benefits. So, other parameters are involved in fitness function with respect to this issue. The way of parameter's relation with fitness function should be determined. For example, in the problem that the time reduction is surveyed, whatever the time be shorter, the features are more suitable.

So, the cost parameter is placed in the denominator of fitness function.

7 Conclusion & Future Work

Researchers should use some techniques for reduction of dimensions in order to analyze them because of large volume of information.

In this way, feature selection methods reduce the effective number features and perform analysis in an effective form. Genetic algorithm is one the more efficient and more applicable feature selection methods in this field. It has a lot of advantages in most of fields such as high speed of algorithm, processing of different modes, acting well even with little background knowledge.

In this field, future researches can be done on combination of genetic algorithm with other algorithms such as phase algorithms in different issues for effective feature selection.

Refrence

[1] H. L. Wei and S. A. Billings, "Feature subset selection and ranking for data dimensionality reduction," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 29, pp. 162-166, 2007.

[2] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," Expert Systems with Applications, vol. 36, pp. 7675-7680, 2009.

[3] İ. Babaoglu, O. Findik, and E. Ülker, "A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine," Expert Systems with Applications, vol. 37, pp. 3177-3183, 2010.

[4] R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, A. Ghandeharioun, R. Boghrati, et al., "diagnosis of coronary arteries stenosis using data mining," Journal of medical signals and sensors, vol. 2, pp. 153-159, 2012.

[5] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial intelligence, vol. 97, pp. 273-324, 1997

[6] M. Dash and H. Liu, "Feature selection for classification," Intelligent data analysis, vol. 1, pp. 131-156, 1997

[7] H. Liu, L. Yu, "Toward integrating feature selection algorithms for classification and clustering," Knowledge and Data Engineering, IEEE Transactions on, vol. 17, pp. 491-502, 2005

[8] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," Machine learning, vol. 3, pp. 95-99, 1988.

[9] J. L. Ribeiro Filho, P. C. Treleaven, and C. Alippi, "Genetic-algorithm programming environments," Computer, vol. 27, pp. 28-43, 1994.

[10] I. Turkoglu, A. Arslan, and E. Ilkay, "An expert system for diagnosis of the heart valve diseases," Expert systems with applications, vol. 23, pp. 229-236, 2002.

[11] A. Damtew, "school of graduate studies school of information science and school of public health," B. Sc. Thesis, Addis Ababa University, 2011.

[12] L, H. Gyu, K. Y. Noh,K. H. Ryu. "A data mining approach for coronary heart disease prediction using HRV features and carotid arterial wall thickness," BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on. Vol. 1. IEEE, 2008.

[13] H. Yan, J. Zheng, Y. Jiang, C. Peng, and S. Xiao, "Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm," Applied Soft Computing, vol. 8, pp. 1105-1111, 2008.

[14] K. Man, K. Tang, and S. Kwong, "Genetic algorithms: concepts and applications [in engineering design]," Industrial Electronics, IEEE Transactions on, vol. 43, pp. 519-534, 1996.

[15] L. Min and W. Cheng, "A genetic algorithm for minimizing the makespan in the case of scheduling identical parallel machines," Artificial Intelligence in Engineering, vol. 13, pp. 399-403, 1999.

[16] H. G. Lee, K. Y. Noh, and K. H. Ryu, "A data mining approach for coronary heart disease prediction using HRV features and carotid arterial wall thickness," International Conference on BioMedical Engineering and Informatics, 2008. BMEI 2008., 2008, pp. 200-206.

[17] Z. Michalewicz, Genetic algorithms+ data structures= evolution programs vol. 3: springer, 1996.

[18] T. Blickle and L. Thiele, "A Mathematical Analysis of Tournament Selection," in ICGA, 1995, pp. 9-16.