A graph-based feature selection method for improving medical diagnosis

A. R. Noruzi
[1] Department of Computer Science, Ashtian Branch, Islamic Azad University, Ashtian, Iran.

H. R. Sahebi
[2] Department of Mathematics, Ashtian Branch, Islamic Azad University, Ashtian, Iran.
sahebi@aiau.ac.ir

Abstract

Classification systems have been widely utilized in medical domain to explore patient's data and extract a predictive model. This model helps physicians to improve their prognosis, diagnosis or treatment planning procedures. Models based on data mining and machine learning techniques have been developed to detect the disease early or assist in clinical breast cancer diagnoses. Medical datasets are often classified by a large number of disease measurements and a relatively small number of patient records. All these measurements (features) are not important or irrelevant/noisy. Feature selection is commonly applied to improve the performance of models. Feature selection is one of the most common and critical tasks in database classification. It reduces the computational cost by removing insignificant features. Feature selection methods can help select the most distinguishing feature sets for classifying different cancers. Consequently, this makes the diagnosis process accurate and comprehensible. This paper presents a graph based feature selection method for medical database classification. Sex benchmarked datasets, which are available in the UCI Machine Learning Repository, have been used in this work. The classification accuracy shows that the proposed method is capable of producing good results with fewer features than the original datasets.

Keywords: *Feature selection, medical dataset, Graph clustering, Feature clustering.*

1. Introduction

The revolution in database technologies has resulted in an increase of data accumulation in many areas, such as financial, marketing and the biological and medical sciences. It has become crucial to locate hidden information by scrutinizing these data effectively. Data mining techniques have been discussed widely and applied successfully in the areas of medical research, scientific analysis and business applications. Recently, the absorption of data mining techniques in medical diagnosis has provided new insights in a large number of medical applications. Feature selection has many advantages such as shortening the number of measurements, reducing the execution time and improving transparency and compactness of the suggested diagnosis [1] [2].

Data mining and machine learning techniques have been used to analyze breast cancer diagnoses, and they have been used to create models to detect the disease early or assist the diagnosis understanding. Because many features are noisy and redundant, especially in high-dimensional data representations, the created models usually suffer from noisy features participating in the training process and then compromise a satisfactory performance. For example, in classification (or clustering) learning algorithms, biased classifiers (or clusters) obtained using noisy and redundant features in the forecasting (or partitioning) process are not reliable.[3] [4].

Feature selection plays an important role in the world of machine learning and more specifically in the classification task. On the one hand the computational cost is reduced and on the other hand, the model is constructed from the simplified data and this improves the general abilities of classifiers. The first motivation is clear, since the computation time to build models is lower with a smaller number of features. The second reason indicates that when the dimension is small, the risk of ‘‘overfitting’’ is reduced [5].

The feature selection methods can be classified into four categories including filter, wrapper, hybrid and embedded models [5-7]. The filter approach relies on the characteristics of the learning data and selects a subset of features without involving any learning model. In contrast, the wrapper approach requires one predetermined learning model and selects features with the aim of improving the generalization performance of that particular learning model. Although the wrapper approach is computationally expensive than the filter approach, the generalization performance of the former approach is better than the later approach. The hybrid approach attempts to take advantage of the filter and wrapper approaches by exploiting their complementary strengths. Embedded methods are embedded in and specific to a given machine learning algorithm, and select the features through the process of generating the classifier.

A feature selection method may be evaluated according to efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of subset of features. These issues are in conflict with each other, generally improving one of them causes reducing the other one. In other words, the filter-based feature selection methods have been paid much attention to the computational time and typically are faster, while the unsupervised wrapper methods usually consider the quality of selected features. Therefore, a trade-off between these two issues has become an important and necessary goal to providing a good search method. Keeping these in mind, in this paper we propose a novel graph-based feature selection method by integrating the concept of graph clustering with the fisher score.

The rest of the paper is organized as follows: Section 2 gives a brief review of previous works. Section 3 presents the proposed feature selection method based on graph theoretic approach. Section 4 reports the experimental results on well-known medical dataset. Finally, Section 5 presents the conclusion.

2. Related Work

Feature selection has been a fertile field of research and development since 1970s in statistical pattern recognition, machine learning, data mining, and there have been a number of attempts to review the feature selection methods [6, 8, 9]. In this section, we briefly review various feature selection methods that can be classified into four categories including filter, wrapper, embedded, and hybrid approaches. Moreover, graph based feature selection methods are also reviewed.

According to whether the class labels of training data are available, feature selection algorithms can be roughly grouped into two families, i.e., supervised feature selection and unsupervised feature selection. Generally speaking, supervised feature selection usually yields better and more reliable performance, mainly because of the utilization of class labels. Given sufficient labeled data, it is possible for supervised algorithms to train appropriate feature selection functions. However, labeling a large number of training data is tedious and time-consuming. In many real world applications, the performance of the existing feature selection algorithms is usually restrained by the paucity of labeled training data. Therefore, it turns out to be a great research challenge to design a feature selection algorithm for the cases when only a few labeled data per task are available [10].

Recently, a number of researchers have focused on several feature selection methods and most of them have reported their good performance in database classification.

Lin et al. [11] applied a Particle Swarm Optimization-based approach to search for appropriate parameter values for a back propagation network to select the most valuable subset of features to improve classification accuracy. Unler et al. [12] developed a modified discrete particle swarm optimization algorithm for the feature selection problem and compared it with tabu and scatter search algorithms to demonstrate its effectiveness.

Chang et al. [13] introduced a hybrid model for integrating a case-based reasoning approach with a particle swarm optimization model for feature subset selection in medical database classification. Salamo et al. [14] evaluated a number of measures for estimating feature relevance based on rough set theory and also proposed three strategies for feature selection in a Case Based Reasoning classifier. Qasem et al. [15] applied a time variant multi-objective particle swarm optimization to an RBF Network for diagnosing medical diseases.

In [16], a new supervised feature selection methods based on hybridization of Particle Swarm Optimization (PSO), PSO based Relative Reduct (PSO-RR) and PSO based Quick Reduct (PSO-QR) are presented for the diseases diagnosis.

Chen [17] aims to present a hybrid intelligence model that uses the cluster analysis techniques with feature selection for analyzing clinical breast cancer diagnoses. Our model provides an option of selecting a subset of salient features for performing clustering and comprehensively considers the use of most existing models that use all the features to perform clustering.

In [18], the K-SVM based on the recognized feature patterns has been proposed. It can be competitively compared with traditional data mining methods in cancer diagnosis. For the phase of feature extraction, the traditional methods of extracting.

3. Proposed method

Recently, the graph-based methods, such as spectral embedding[19], spectral clustering [20], and semi-supervised learning [21], have played an important role in machine learning due to their ability to encode similarity relationships among data. In feature selection, by representing the feature space into a graph, the graph based methods can provide a universal and flexible framework that reflects underlying manifold structure and relationships between feature vectors.

In this section the feature clustering of the search space, graph clustering and select best representative feature from each cluster are described.

3.1 Graph representation

A preliminary step for all graph-based methods is to represent training data with an undirected graph. For this purpose, the feature set is mapped into its equivalent graph $G = (F, E, w_F)$, where $F = \{F_1, F_2, \dots, F_n\}$ is a set of original features, $E = \{(F_i, F_j): F_i, F_j \in F\}$ denotes the edges of graph and w_{ij} indicates similarity between two features F_i and F_j connected by the edge (F_i, F_j). Different measures for computing vertex similarities (i.e. edge weights) leads to different performances on the graph-based feature selection methods. In this work, we have used well-known Pearson product-moment correlation coefficient [22] to measure similarity between different features of a given training set.

3.2 Feature clustering

Feature clustering is an efficient approach for dimensionality reduction [23, 24]. The main idea of feature clustering is to group original features into different clusters based on their similarities; thus, the features in the same clusters are similar to each other. Quite different from existing feature clustering algorithms, in this paper a community detection method is applied to cluster the features into different groups. The community structure is one of the most important patterns in network. Since finding the communities in the network can significantly improve our understanding of the complex relations, lots of work has been done in recent years [25, 26]. In this work, we have used the Louvain community detection algorithm [27] to identify the feature clusters. This algorithm detects communities in the graph by maximizing a specific modularity function. This method has two advantages. First, its steps are intuitive and easy to implement, and second, the algorithm is extremely fast.

3.3 Select representative feature

The main purpose of this step is to identify relevant and influential features from each cluster. In other words, in each cluster, some of high relevance features are retained and the others will be removed. To this end, fisher score [28] is utilized to identify representative features. After calculating the efficient value of features, some feature with efficient value less than δ parameter are removed and reminder feature are select as final feature set.

4. Experimental results

The classification performance of the proposed feature selection method is measured using an SVM classifier. The performance of the proposed method is evaluated using five benchmark datasets: Wisconsin Breast Cancer, Pima Indians Diabetes, Heart-Statlog, Hepatitis and Cleveland Heart Disease, which are available from the UCI Machine Learning Repository. Table 1 summarizes the number of features, instances and classes for each dataset used in this study.

Table 1: Details of used datasets

Dataset	Features	Samples	Classes
Wisconsin Breast Cancer	9	699	2
Pima Indians Diabetes	8	768	2
Heart-Statlog	13	270	2
Hepatitis	19	155	2
Cleveland Heart Disease	13	296	5

All datasets are split into 10 subsets of approximately equal size. Randomly, one dataset is used for testing and the remainder are used for training. The same procedure is repeated 10 times and the mean classification accuracy is computed. Tables 2 presents the results reported for each dataset. The proposed method was compared to the well-known filter-based methods, including, Fisher Score (FS) [28], mRMR [29] and n feature. From the results it can be observed that in most cases the proposed method obtained the highest classification accuracy compared to those of filter-based methods. For example, for the Hepatitis dataset, proposed method obtained an 81.93% classification accuracy while for FS and mRMR this value was reported 80.73 and 82.73 correspondingly.

Table 2: Classification results with different methods

Dataset	Feature selection method			
	FS	mRMR	Proposed method	All features
Wisconsin Breast Cancer	94.81	95.91	96.42	95.85
Pima Indians Diabetes	71.64	70.78	75.46	73.63
Heart-Statlog	82.39	83.41	84.91	84.36
Hepatitis	80.73	82.73	82.32	81.93
Cleveland Heart Disease	82.69	83.54	85.68	83.27

Moreover, additional experiments were conducted to compare the proposed method with the other feature selection method based on the different number of selected features. Figs. 1 and 2 plot the classification accuracy (average over 10 independent runs) curves of SVM classifiers on Wisconsin Breast Cancer and Heart-Statlog, respectively. In all the plots, the x-axis denotes the subset of selected features, while the y-axis is the average classification accuracy. Fig. 1 shows that the proposed method is superior to the other methods applied on the SVM classifier.

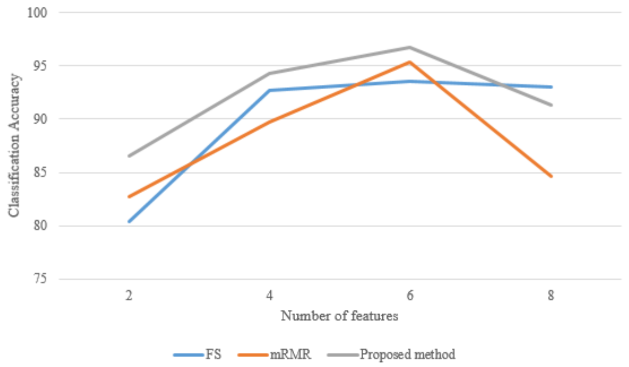

Fig 1: classification accuracy with number of features with different methods on Wisconsin Breast Cancer

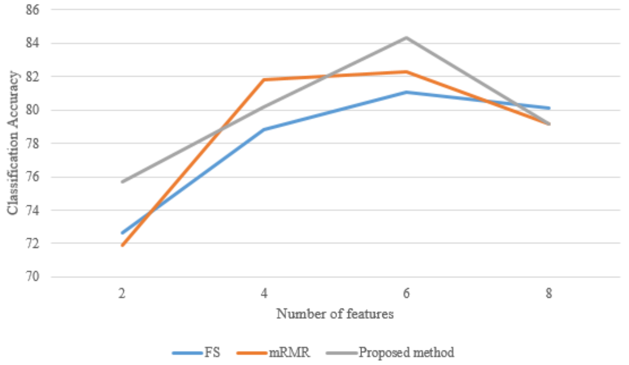

Fig 2: classification accuracy with number of features with different methods on Heart-Statlog

5. Conclusions

Identifying key biomarkers for different cancer types can improve diagnosis accuracy and treatment. Gene expression data can help differentiate between cancer subtypes. However the limitation of having a small number of samples versus a larger number of genes represented in a dataset leads to the over fitting of classification models. Feature selection aims to reduce the amount of unnecessary, irrelevant and redundant features. It helps retrieve the most relevant features in datasets and improves the classification accuracy with less computational effort. If the features are not chosen well, even the best classifier performs poorly. In this paper, we describe a graph-based feature selection method with an SVM classifier. The intention is to select the correct set of features for classification when datasets contain noisy, redundant and vague information.

The proposed methods are compared with well-known filter-based feature selection method and classification accuracy measures are used to evaluate the performance of the proposed approaches. Hence the analysis section clearly proved the effectiveness of proposed method for diagnosis the disease over the other existing approaches.

References

[1]. R.E. Abdel-Aal, GMDH based feature ranking and selection for improved classification of medical data, J. Biomed. Inform. 38(6) (2005) 456–468.

[2]. M. F. Akay, Support vector machines combined with feature selection for breast cancer diagnosis, Int.J. Expert Syst. Appl. 36(2) (2009) 3240–3247.

[3]. H. Liu, L. Yu, Toward integrating feature selection algorithms for classificationand clustering, IEEE Transactions on Knowledge and Data Engineering 17 (4)(2005) 491–502.

[4]. M.A. Jayaram, A.G. Karegowda, A.S. Manjunath, Feature subset selection prob-lem using wrapper approach in supervised learning, International Journal ofComputer Applications 1 (7) (2010) 13–16.

[5]. 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.

[6]. Saeys, Y., I. Inza, and P. Larranaga, A review of feature selection techniques in bioinformatics. Bioinformatics, 2007. **23**(19): p. 2507-17.

[7]. Song, Q., J. Ni, and G. Wang, A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2013. **25**(1): p. 1 - 14.

[8]. Chandrashekar, G. and F. Sahin, A survey on feature selection methods. Computers & Electrical Engineering, 2014. **40**(1): p. 16-28.

[9]. Liu, H. and L. Yu, Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2005. **17**(4): p. 491 - 502

[10]. Yi Yang, et al., Feature Selection for Multimedia Analysis by Sharing Information Among Multiple Tasks. Multimedia, IEEE Transactions on 2012. **15**(3): p. 661 - 669

[11]. Shih-WeiLin, Shih-Chieh Chen, Wen-Jie Wu, Chih-Hsien Chen, Parameter determination and feature selection for back-propagation network by particle swarm optimization, Int.J.Knowl.Inf.Syst. 21(2) (2009) 249–266.

[12]. Alper Unler, Alper Murat, A discrete particle swarm optimization method for feature selectionin binary classification problems, Eur.J.Oper.Res.206(3) (2010)528–539.

[13]. Pei-Chann Chang, Jyun-JieLin, Chen-Hao Liu, An attribute weight assignment and particle swarm optimization algorithm for medical database classification, Int. J.Comput. Methods Progr. Biomed. 107(3) (2012) 382–392.

[14]. Maria Salamo, Maite Lopez-Sanchez, Rough set based approaches to feature selection for case-based reasoning classifiers, Int.J.Pattern Recognit.Lett.32 (2) (2011) 280–292.

[15]. Sultan Noman Qasem, Siti Mariyam Shamsuddin, Radial basis function network based on time variant multi-objective particle swarm optimization for medical diseases diagnosis, Int.J. Appl. Soft Comput. 11(1) (2011)) 1427–1438.

[16]. 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.

[17]. Chen, C.-H., A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. Applied Soft Computing, 2014. **20**(0): p. 4-14.

[18]. Zheng, B., S.W. Yoon, and S.S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 2014. **41**(4, Part 1): p. 1476-1482.

[19]. Belkin, M. and P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering. Neural Inform. Process. Systems 1, 2002: p. 585-592.

[20]. Shi, J. and J. Malik, Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Machine Intell, 2000. **22**(8): p. 888–905.

[21]. Chung, F., Spectral Graph Theory. In: Regional Conference Series in Mathematics American Mathematical Society, 1997. **92**(92): p. 1-212.

[22]. 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.

[23]. Jung-Yi Jiang, Ren-Jia Liou, and S.-J. Lee, A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification. IEEE Transactions Knowledge and Data Engineering, 2011. **23**(3): p. 335 - 349

[24]. Xi Zhao, W. Deng., and Y. Sh, Feature Selection with Attributes Clustering by Maximal Information Coefficient Procedia Computer Science, 2013. **17**(Pages 70–79).

[25]. Chuan Shi, et al., A link clustering based overlapping community detection algorithm. Data & Knowledge Engineering, 2013. **87**: p. Pages 394–404.

[26]. Yakun Li , et al., Efficient community detection with additive constrains on large networks. Knowledge-Based Systems, 2013. **52**: p. Pages 268–278.

[27]. V. Blondel, J.G., R. Lambiotte, and E. Lefebvre, Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008. **10008**: p. pp. 1–12.

[28]. Quanquan Gu, Zhenhui Li, and J. Han, Generalized Fisher Score for Feature Selection. In: Proceedings of the International Conference on Uncertainty in Artificial Intelligence, 2011.

[29]. Hanchuan Peng, Fuhui Long, and C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions Pattern Analysis and Machine Intelligence, 2005. **27**(8): p. 1226 - 1238