2010 3rd International Conference on Computer and Electrical Engineering (ICCEE 2010)
IPCSIT vol. 53 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V53.No.1.08

A Wrapper-based Feature Selection for Analysis of Large Data Sets

Jinsong Leng[+], Craig Valli and Leisa Armstrong

School of Computer and Security, Edith Cowan University,WA, Australia

Abstract. Knowledge discovery from large data sets using classic data mining techniques has been proved to be difficult due to large size in both dimension and samples. In real applications, data sets often consist of many noisy, redundant, and irrelevant features, resulting in degrading the classification accuracy and increasing the complexity exponentially. Due to the inherent nature, the analysis of the quality of data sets is difficult and very limited approaches about this issue can be found in the literature. This paper presents a novel method to investigate the quality and structure of data sets, i.e., how to analyze whether there are noisy and irrelevant features embedded in data sets. In doing so, a wrapper-based feature selection method using genetic algorithm and an external classifier are employed for selecting the discriminative features. The importance of features are ranked in terms of their frequency appeared in the selected chromosomes. The effectiveness of proposed idea has been investigated and discussed with some sample data sets.

Keywords: Data Mining, Feature Selection, Genetic Algorithm, Classification

1. Introduction

Mining interesting patterns and anomaly trajectories from large data sets has attracted much interests in data mining society. The main purpose of knowledge discovery is to group similar expression patterns or profiles over a large of percentage of objects in data sets. Due to the inherent nature of data sets, each data mining technique only imposes a certain structure on the data, thereby resulting in no one-size-fits-all mining methods available for a variety of practical applications. Practically, many real-world applications cannot directly be applied to data mining algorithms due to many noisy, redundant, and irrelevant features hidden in data sets. Meanwhile, there are some additional difficulties of knowledge engineering in large data sets: One is the problem so-called 'the curse of dimensionality' [2]; the other is the adaption of mining algorithms with the data sets changing over the time. The former involves the dimensionality reduction, while the latter deals with the data streams with dynamic changes.

As we know, the higher ratio of the number of training sample to the number of dimensions (features) used by classifier, the better generation ability of the resulting classifier. The presence of noise and irrelevant features makes accurate classification very difficult. For instance, microarray data usually consists of thousands of features (genes) with only few dozen of samples [14]. Obviously, selecting a relevant subset of features is of paramount importance in mining large high dimensional data sets. In addition, the complexity grows exponentially with the increase of the number of features, making the search of all possible spaces infeasible.

Traditional classification/clustering methods try to group clusters in full-dimensional space with either distance-based measure or density-based measure, including k-means clustering, self-organizing maps (SOM), and hierarchical methods, and so on [5]. Normally, the large data sets such as microarray data are much larger in dimensionality than some sample data sets used in conventional data mining algorithms. Moreover, they consist of much noise, redundant, and irrelevant attributes due to the natural variation, and high internal de-

[+] Corresponding author.
E-mail address:j.leng@ecu.edu.au.

pendences between features in real applications. Selecting fewer discriminative attributes by removing irrelevant and/or redundant attributes can not only increase the classification accuracy but also reduce the computational complexity to alleviate 'the curse of dimensionality' problem. In this respect, the investigation of the quality and structure of data sets is of paramount importance to the success of data mining algorithms.

Feature selection is a necessary step for most of real applications. For example, a feature selection approach in [12] has been applied to the evaluation of data attributes with regard to their utility for wheat Yield Prediction. The forward feature selection is considered as the complete search strategy, wherein support vector machine and RegTree are used for classification. Another filter based feature selection approach [15] has been used for the satellite images by evaluating and selecting the features in terms of the related covariance matrices.

The purpose of feature selection is to find the 'good' features by just selecting one representative feature subset from all features. Inadequate removal of attributes may result in the massive losses of internal information among features. Even though a few features can lead to a good classification accuracy, the additional features being added may not contribute much to the performance but they would not degrade the overall performance. In such case, the data set is well structured and behaved. Apparently, it would be beneficial to select as many as possible features from large data sets to discovery more information among those features. If the performance degrades, the data set contains noisy, irrelevant features and thus is not well constructed. Consequently, the analysis of the noisy and irrelevant features can also shed lights for investigating the structure of data sets.

The focus of this paper is on the analysis of quality of data sets by identifying the eliminating noisy, redundant, and irrelevant features embedded in data sets. A novel idea has been proposed to investigate the internal dependences between features and identify their relative importance of features. In doing so, a wrapper-based feature selection algorithm with genetic algorithm (GA) and K nearest neighbor (KNN) has been developed to rank the importance of features. In this paper, we discuss how the noise and irrelevant features in data sets can be identified. In addition, the problem of selecting minimum discriminative is also analyzed and specified.

The rest of this paper is organized as follows: Section II gives an overview on feature selection methods. The details of GA/KNN algorithm are described in Section III. The empirical studies and discussions are presented, and the analysis of noisy and irrelevant features is discussed in Section IV. Finally, Section V concludes the paper.

2. Feature Selection

Knowledge mining involves the following steps: data acquisition, feature selection, and classification or clustering. As described above, the purpose of feature selection is to reduce the complexity and enhance the classification accuracy. Despite of the size of data sets, it may be beneficial to remove the noisy, redundant, and irrelevant features from raw data sets before applying data mining algorithms.

Feature selection is the measure process to find the subset of features in terms of the importance of features, whereby the importance of feature can be estimated by some mathematical and clustering based measure criteria. According to the way of computing the feature evaluation indices, feature selection can be broadly classified into three categories: filter approach, wrapper approach, and hybrid approach [10].

The filter approach computes the feature evaluation weight but without performing classification of data, eventually finding the 'good' subset of features. Most measure criteria for evaluating features are actually statistics based in nature. The principle of filter approaches is to select the subset of features which have high dependency on target class and while have less correlation among them. One group of filter methods is to measure the importance by maximizing the clustering performance. Other approaches are to find redundant or irrelevant feature to be removed that carries little or additional information using statistics measures [7, 16]. Until now, a number of filter methods have been proposed for feature selection, including sequential forward search [4], sequential floating forward search [11], stepwise clustering [7], feature selection through clustering[10] and so on.

Differently, the wrapper-based methods employ some inductive classification algorithm(s) to evaluate the goodness of subset of features being selected. The performance of this approach relies on two factors: the strategy to search the space of all possible feature subsets; and the criterion to evaluate the classification accuracy of the selected subset of features. The purpose of feature selection is to remove noisy, redundancy, and irrelevant features from raw data sets while minimizing the information loss. Finding the minimum subset of features is quite difficult in large data sets. As reported [8], the wrapper-based approaches can perform better than filter based approaches. Some hybrid approaches have also been proposed in conjunction with some filters and the inductive classification algorithms [17].

The wrapper approaches of feature selection aim to find the minimum discriminative features to reach the high classification accuracy, while the filer approaches are to compute the 'best' subset of features in terms of some criteria.However, the inherent nature among features such as function regulation and frequent patterns [3] has been ignored in both filter and wrapper approaches. The major disadvantage of those methods is that each subset of features is evaluated regarding their dependencies, thereby ignoring the function regulation among features [13].

Rather than focusing on the classification accuracy of selected subset of features, this paper uses the feature selection for the analysis of the quality of data sets, i.e., whether the data sets contain much noisy and redundant features. The wrapper approach, a genetic algorithm in conjunction with K nearest neighbors (GA/KNN), is employed for selecting the discriminative features. The selected features are validated by various classification algorithms, in order to derive the effective approach for identifying noisy, redundant, and irrelevant in data sets.

3. Wrapper based Feature Selection

3.1 Genetic Algorithm

Genetic algorithms (GAs) was originally introduced by John Holland in 1975 [6]. GA is based on the random sampling strategy to find suboptimal solutions. Due to its stochastic nature, GAs can be over fitting and risky falling in the local optimum. However, the studies shows that Gas have some advantages over other filter methods and heuristic search strategies [8]. GA algorithms have been widely used to alleviate the high dimensional problems, such as microarray data [9, 17].

GA generates the sample population with certain number of chromosomes. Here a chromosome is a subset of features randomly selected from the data set. The 'goodness' of each chromosome is evaluated in terms of the fitness, which is estimated by a classification algorithm. The certain number of chromosomes (subset of features) that meet the fitness criterion can be found using an iterative scheme. For each iteration, the fitness of each chromosome is computed. The chromosome is chosen for crossover and mutation based on the 'roulette wheel selection' strategy, which assigns the higher probability to select the chromosome with better fitness, and versa. The purpose of the use of the 'roulette wheel selection' is to give the more chance for 'good' chromosome being selected so as to evolve better next generation. A new offspring (chromosome) is generated by making the crossover between two selected chromosomes. After a crossover is performed, mutation is employed. Mutation changes randomly the new offspring in order to prevent falling all solutions in population into a local optimum. If the fitness of the chromosome is satisfied, one near optimal chromosome (subset of features) is obtained, and a new population is generated. The computation is terminated until the predefined K near optimal chromosomes is found.

The basic steps of GAs are indicated as follows:

1. Population: Generate random population of n chromosomes (subset of features).
2. Fitness: Evaluate the related fitness of each chromosome.
3. Iteration: Repeat until the predefined number N of chromosomes found:
   a) Selection: Choose two chromosomes for crossover.
   b) Crossover: Form a new chromosomes based on the crossover strategy.
   c) Mutation: Mute the new chromosomes by the mutation probability.
   d) Fitness: Compute the fitness of the muted chromosome.
   e) Update: Replace the muted chromosome in the population.

f) Evaluation: If fitness is satisfied:
   i. Keep this chromosome.
   ii. Generate a new population, and compute the fitness of each chromosome.
4. Return: Find N chromosomes (subset of features).

3.2 K nearest neighbor

K nearest neighbor (KNN) algorithm is a non-parameterized supervised learning algorithm. KNN finds the number of K neighbors with minimum distance to the query instance in the training samples. The major advantage of KNN is simple and model-free. Given a data sample, the K number of training samples closest to the data sample can be found, and then the given data sample is classified by the K nearest neighbors using the consensus or majority rule. If a data sample is correctly classified, it is assigned a scoring 1, otherwise 0. The fitness of a chromosome is the summation of the scoring of all training samples with the subset of features. If the fitness is satisfied, a near optimal subset of features is thus found.

4. Empirical Studies

4.1 Sample data sets

Normally, the size of the data sets can be categorized in terms of its dimensionality: low-dimensional ($D \leq 15$); medium-dimensional ($15 < D < 60$); high-dimensional ($D \geq 60$). To validate the proposed idea, six sample real data sets with different dimensionality are chosen from the UCI Machine Learning Repository [1]. The details of six data sets are described as follows:

1. Wine. The data set contains 13 continuous features and 178 instance with three classes: class 1 (59), class 2 (71), and class 3 (48).
2. Wisconsin Diagnostic Breast Cancer (WDBC). WDBC has 30 real-valued features and 569 instances with two classes: benign (357), malignant (212).
3. Ionosphere. This data set consists of 34 continuous features and 351 instances with two classes: good (225) , bad (126).
4. Connectionist Bench (Sonar). Sonar data set contains 60 features and 208 samples with two classes: rock (97), mine (111).
5. Hill Vally. The data set consists of 240 instances and 100 noisy features, which is split from the hill vally with noisy data set.
6. MUSK ”Clean1”. This data set has 168 features and 476 instances with two classes: musk (207), non-musk (269).

4.2 Parameters of GA/KNN

The GA/KNN algorithm is implemented in C# and . NET. Each data set is randomly split into training data set and testing data. The population size is set to 100. The features in each chromosome is 10. The crossover is not used in this experiment due to the duplicate features in two chromosome. The different mutation rate is considered: between 1 and 5 of its genes are randomly selected for mutation. The number of mutations (from 1 to 5) is assigned randomly, with probabilities, 0.53125, 0.25, 0.125, 0.0625, and 0.03125, respectively [9]. The criterion of good fitness for a chromosome is considered as 80% correctly classification. As suggested in [9], the features are ranked in terms of the frequency in the selected chromosomes. After that, we can obtain the classification accuracy over the testing data set using different number of the top ranked features. Based on the classification accuracy, we can easily to analyze and specify the noisy and irrelevant in the original data set.

4.3 Experimental results

The original data set is divided into training data set and testing data set using randomly sampling strategy. The three independent runs GA/KNN over each training data set are conducted. Each separate run generates 3000 chromosomes (subset of features). The features in each data set are ranked in terms of the frequency selected in 3000 chromosomes. Then the testing results are computed using the testing data set with the different number of ranked features.

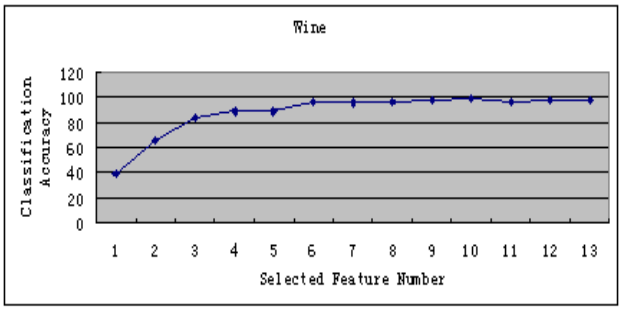

Fig. 1. Wine: Number of Features and Classification Accuracy

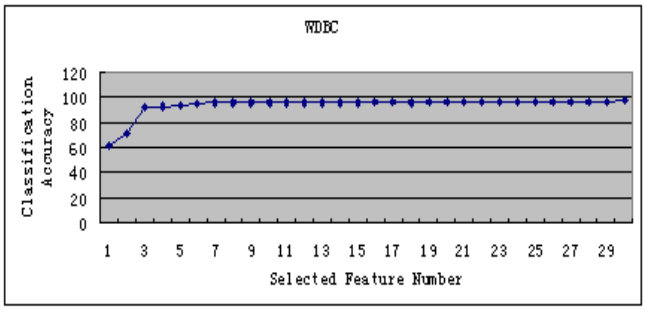

Fig. 2. WDBC: Number of Features and Classification Accuracy

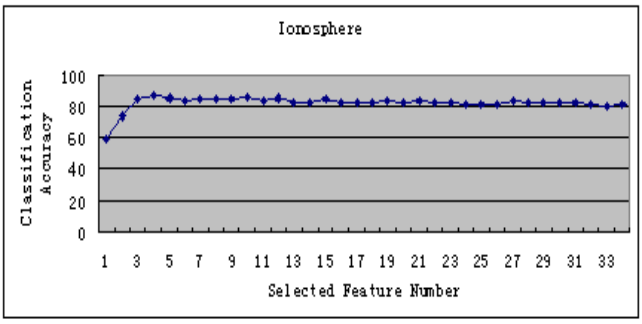

Fig. 3. Inosphere: Number of Features and Classification Accuracy

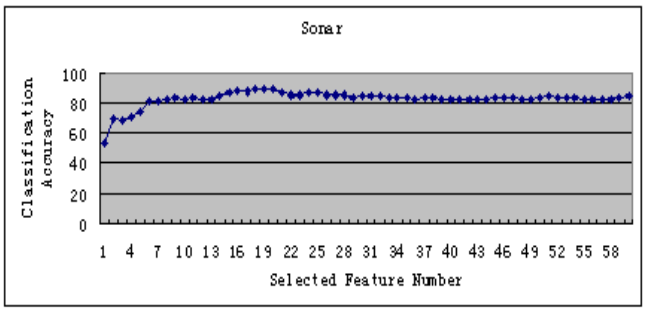

Fig. 4. Sonar: Number of Features and Classification Accuracy

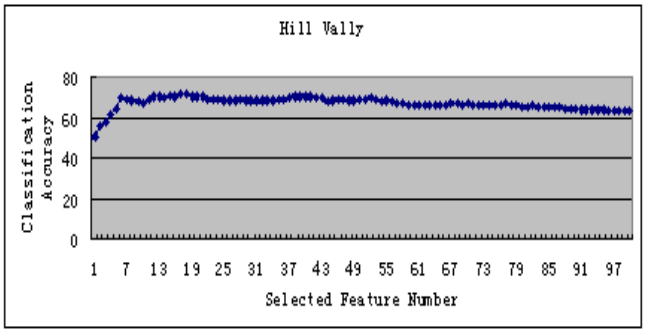

Fig. 5. Hill Vally (Noisy): Number of Features and Classification Accuracy

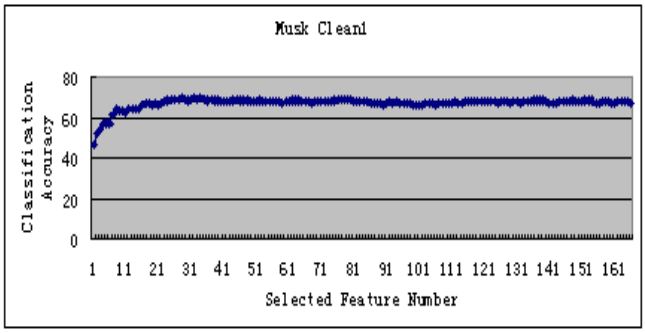

Fig. 6. Musk: Number of Features and Classification Accuracy

For the six data sets, the relations between the number of selected features and classification accuracy are illustrated in Figure 1 ~ 6, respectively.

4.4 Discussion and interpretation

All data sets except the Hill Vally have been extensively studies using various classifiers [1]. As indicated in Section 4.3, it is clear that we can easily get the high classification accuracy with very fewer features. With the increase of the number of features being considered, the classification accuracy is quite steady, indicating that they have 'well behaved' class structures without much noisy and irrelevant features being included. As shown in Fig. 1, the increase of number of features in wine data set will improve the performance slightly, indicating that there exists a tight relationship among features. The similar relations can also be found in WDBC, Sonar Inosphere, and Musk Clean 1 data sets, even some few features might be the redundant and irrelevant features hidden in those data sets. Overall, those data sets are well structured. However, the classification accuracy of the Hill Vally data set degrades significantly from 70.8% to 62.9%, as the number of features increases. This is good indicator that the lower ranked features may consist of many noisy features in this data set. We can find that first 50 features selected perform better, while the performance degrades with the increase of number of selected features. In such case, the last fifty features are suspicious and require the further identification whether they are noisy and irrelevant features in the Hill Vally (noisy) data sets.

With the analysis of performance, we can choose the subset of features that have strong inherent relationship for classification or clustering using different data mining algorithms. Furthermore, we can get the clear picture about the structure of data and detect the noisy and irrelevant features with the low ranking scores.

5. Conclusion

It is of great of importance to remove the noisy and irrelevant features and data samples embedded in data sets before applying some data mining techniques to analyze the data sets. This paper describes a novel idea to identify the noisy and irrelevant features embedded in data sets and detect the quality of the structure of data sets. Conventional approaches of the use of GA/KNN are intended to find the minimum discriminative subset of features according to the classification accuracy. Finding the discriminative subset of features will lead to the losses of lots of useful information such as frequent patterns and regular functions among features. This paper uses the GA/KNN to evaluate the quality of data sets in order to remove the noisy features in original data sets. The analysis on data structure and removal of noisy and irrelevant features in large data sets can result in the high and steady performance for various classifiers.

This paper proposes a novel idea to investigate the structure of the data set and reveal the inherent relationship among features. The ultimate goal of this research is to the discriminative features with good frequent patterns, in order to reveal the regular functions hidden in the features. This research may lead to a better solution to many practical problems with respect to applications to agriculture and bioinformatics.

6. References

[1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[2] R. Bellman. Dynamic Programming. Princeton University Press, Princeton, NJ, 1957.

[3] Y. Cheng and G. M. Church. Biclustering of Expression Data. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pages 93–103, 2000.

[4] P. A. Devijver and J. Kittler. Pattern Recognition: A Statistical Approach. Englewood Cliffs: Prentice Hall, 1982.

[5] X. W. et. al. Top 10 algorithms in data mining. Knowl. Inf. Syst., 14(1):1–37, 2007.

[6] J. H. Holland. Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor, 1975.

[7] B. King. Step-Wise Clustering Procedures. Journal of the *Amer*ican Statistical Association, 62(317):86–101, 1967.

[8] M. Kudo and J. Sklansky. Comparison of Algorithms that Select Features for Pattern Classifiers. Pattern Recognition, 33(1):25–41, 2000.

[9] L. Li, C. Weinberg, T.A.Darden, and L. Pedersen. Gene Selection for Sample Classification based on Gene Expression Data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics, 17(12):1131–1142, 2001.

[10] P.Mitra, C.Murthy, and S. Pal. Unsupervised Feature Selection Using Feature Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24:301–312, 2002.

[11] P. Pudil, J. Novovi˘cov´a, and J. Kittler. Floating Search Methods in Feature Selection. Pattern Recogn. Lett., 15(11):1119–1125, 1994.

[12] G. Ruß and R. Kruse. Feature Selection for Wheat Yield Prediction. In Research and Development in Intelligent Systems XXVI, Incorporating Applications and Innovations in Intelligent Systems XVII, pages 465–478, 2010.

[13] Y. Saeys, I. Inza, and P. Larranaga. A Review of Feature Selection Techniques in Bioinformatics. Bioinformatics, 23(19):2507–2517, 2007.

[14] R. L. Somorjai, B. Dolenko, and R. Baumgartner. Class Prediction and Discovery using Gene Microarray and Proteomics Mass Spectroscopy Data: Curses, Caveats, Cautions.Bioinformatics, 19(12):1484–1491, 2003.

[15] C. D. Stefano, F. Fontanella, and C. Marrocco. A GA-based Feature Selection Algorithm for Remote Sensing Images. In EvoWorkshops 2008, pages 285–294, 2008.

[16] S. Thepdpridis and K. Koutroumbas. Pattern Recognition Algorithms for Data Mining. CHAPMAN & HALL/CRC, 2004.

[17] P. Yang, B. Zhou, Z. Zhang, and A. Zomaya. A Multi-filter Enhanced Genetic Ensemble System for Gene Selection and Sample Classification of Microarray Data. BMC Bioinformatics, 11(Suppl 1):S5, 2010.