A New Approach for Wrapper Feature Selection Using Genetic Algorithm for Big Data

Waad Bouaguel

Abstract The increased dimensionality of genomic and proteomic data produced by microarray and mass spectrometry technology makes testing and training of general classification method difficult. Special data analysis is demanded in this case and one of the common ways to handle high dimensionality is identification of the most relevant features in the data. Wrapper feature selection is one of the most common and effective techniques for feature selection. Although efficient, wrapper methods have some limitations due to the fact that their result depends on the search strategy. In theory when a complex search is used, it may take much longer to choose the best subset of features and may be impractical in some cases. Hence we propose a new wrapper feature selection for big data based on a random search using genetic algorithm and prior information. The new approach was tested on 2 biological dataset and compared to two well known wrapper feature selection approaches and results illustrate that our approach gives the best performances.

Keywords Wrapper · Feature selection · Big data

1 Introduction

Over the past few years, the problem of understanding cancer treatment went from basic to one of the most important task in data mining, thanks to expanding knowledge of cancer genomics and the technologies that make such understanding possible [1].

Genomic sequencing is continuously changing the way we understand cancer. Over the time, we have come to a point where the challenge is not so much how to generate large amounts of data, but how to connect the enormous amounts of genomic data churned out by ever-advancing technologies so that they translate into

W. Bouaguel(✉)
LARODEC, ISG, University of Tunis, Tunis, Tunisia
e-mail: bouaguelwaad@mailpost.tn

K. Lavangnananda et al. (eds.), *Intelligent and Evolutionary Systems*,
Proceedings in Adaptation, Learning and Optimization 5,
DOI: 10.1007/978-3-319-27000-5_6

meaningful cancer prevention and treatment strategies. As there are thousands of gene expressions and only a few dozens of observations in a typical gene expression data set, the number of genes d is usually of order 1000 to 10000 while n the number of biological observations is somewhere between 10 and 100 [1]. Such a condition makes the application of many classification methods a hard task.

Feature selection aims at identifying a subset of features for building robust learning models. Since only a small number of genes among tens of thousands show strong correlation with the targeted disease, some works address the problem of defining which is the appropriate number of genes to select [2, 3]. The choice of the best set of pertinent features to retain is a key factor for a successful and effective classification [1]. In general, redundant and irrelevant features can never help to improve the performance of a classifier or a model. However, they are usually added by mistake to the learning process. Let's take the case of cancer diagnosis where we aim to study the link between the symptoms and their class of diseases. For example, If the patient identification (ID) is considered as one of the input features, the classifier my conclude that the class of disease is influenced by the patient ID, which will influence badly the final result. Thus, these kind of features should be removed in order to increase the learning performance.

Usually, a feature selection method try to find a representative subset of features from the original features space. This selected subset should bring the same information of the original feature space and improve the accuracy of a particular application. According to [1] feature selection process may reduce the time complexity of an algorithm and usually facilitate the data understanding.

Feature selection methods can be grouped into two groups: filter and wrapper methods [4]. On one hand, filter methods evaluate features, individually before the learning process and eliminate some. Wrapper methods on the other hand, are an other category of feature selection methods, in which the prediction accuracy of a classifier is used as a threshold to separate the best features from the others. According to [5] wrapper methods generally result in better performance than filter methods because the feature selection process is optimized for the classification algorithm to be used. Typically wrapper approach use some sort of search strategy to generate the candidate subsets. The search strategy is broadly classified as exhaustive (eg. branch & bound), heuristic (eg. forward selection, backward selection), and random search (eg. genetic algorithm (GA)). The search complexity depends of the data dimensionality, it is usually exponential for an exhaustive search and quadratic for a heuristic search and may be linear to the number of iterations for a random search [4]. Hence using random search seems to be to most appropriate choice but the feature space have to be first reduced using some prior information in order to have a linear complexity.

The presence of prior information and additional information about how the features will interact in the classification model have always a great impact on feature selection and on its subsequent application. So whenever possible try to use this information. For example, when the biological relevance of feature can be ascertained, the potentially irrelevant or obvious features can also be eliminated.

Further to enhance the classification accuracy and learning runtime in big data as biological ones, we propose a new wrapper feature selection method that use in a first step prior information to find a minimum set of features in order to reduce the search space then use a random search using genetic algorithm leading to a new set of features such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all features.

This paper is organized as follows. "Wrapper framework" describes the wrapper feature selection approach. "New approach for wrapper feature selection" proposes a two-stage feature selection approach combining prior knowledge and GA. "Experimental investigations" describes the used datasets and the performance metrics. Then, our results are summarized in "Results Analysis" and conclusions are drawn in "Conclusion".

2 Wrapper Framework

Typically a wrapper approach use a generation mechanism to generate candidate subset: The original feature set contains d features, the total number of competing candidate subsets to be generated is 2^d, which is a huge number even for medium-sized d. The ideal feature selection approach is the exhaustive search of the full set of features to find the optimal subset. However, as the number of features increases the exhaustive search becomes rapidly impractical even for a moderate number of features [6]. If we look at different ways in which features subsets are generated among many variations, three basic schemes are available in the literature namely forward selection, backward elimination and random scheme [4].

Forward selection and backward elimination are considered as heuristics. Generally, sequential generation can help in getting a valid subset within a reasonable time but still it cannot find an optimal subset. This is due to the fact that the generation scheme uses a heuristic to obtain an optimal subset by selecting sequentially the best, as in the forward case, or removing the worst as in the backward case. Using such kind of generator will without doubt speed up the selection process. However, if the search falls in a local optima it cannot turn back. In fact the generator has no way to get out of the local optima because what has been removed cannot be added and what has been added cannot be removed. This is a big shortcoming of sequential schemes. To overcome this problem we may use the random generation scheme, to add randomness to the fixed rule of sequential generation and avoid getting stuck at some local optima [7].

Random search works well for search spaces with a high density of good solutions. GA can be considered as a random search algorithm, since randomness is embedded in GA at almost every level [8]. The idea of applying genetic algorithms with wrapper feature selection is not novel. Of these, Yang and Hanovar used genetic algorithm and neural network to investigate feature subset selection [9].

3 New Approach for Wrapper Feature Selection

In this section we propose a novel approach for wrapper feature selection:

- At first, a based on similarity study with the prior knowledge primary dimensionality reduction step is conducted on the original feature space. This step is used to reduce the search space.
- Second, the subset generation step is performed using genetic algorithm.

3.1 Primary Dimensionality Reduction Step: Similarity Study

The first step of our proposed approach is designed specifically to select less redundant features without sacrificing quality. Redundancy is measured by a similarity measure between a preselected set of features and the remaining features in the dataset. In this step we enhance an existing set of preselected features by adding additional features as a complement. In any data mining application we may already have a set of features preselected with prior information. In fact, experts have years of experience on some particular knowledge about which features are more important. This knowledge is generally obtained by years of use of classical feature selection methods. Thus, a possible improvement of any search strategy is to use the prior knowledge and to eliminate redundant features before generating the candidate subsets. Since our goal is to take advantage of any additional information about the feature, we may want to select a set of features complementary to those preselected by experts. Hence, we need to study the effect of using prior information on relevant feature complexity.

First, we split the features set in two sets. The first one regroups a set of features that were assumed to be more relevant according to some prior knowledge. The second set contains the remaining ones. Once the two sets are obtained we conduct a similarity study and a similarity matrix is constructed. In this step the mutual information (MI) is chosen as a similarity measure given its efficiency in providing a solid theoretical framework for measuring the relation between the classes and a feature or more than one feature [10]. Formally, the MI of two continuous random variables X^j and $X^{j'}$ is defined as follows:

$$MI(X^j, X^{j'}) = \int\int p(x^j, x^{j'}) log \frac{p(x^j, x^{j'})}{p(x^j)p(x^{j'})} dx^j dx^{j'}, \tag{1}$$

where $p(x^j, x^{j'})$ is the joint probability density function and $p(x^j)$ and $p(x^{j'})$ are the marginal probability density functions. In the case of discrete random variables, the double integral becomes a summation, where $p(x^j, x^{j'})$ is the joint probability mass function, and $p(x^j)$ and $p(x^{j'})$ are the marginal probability mass functions. MI is an information metric used to measure the relevance of features taking into account the amount of information shared by two features [11]. Large values of MI indicate high correlation between the two features and zero indicates that two features are uncorrelated. Many authors proposed feature selection methods based on MI in different evaluation functions [11, 12].

Finally, we investigate level of similarity of each feature from the remaining set with the features of the first set. If the similarity is over 80%, the evaluated feature is eliminated else it is retained for further examination. More details are given in Fig. 1.

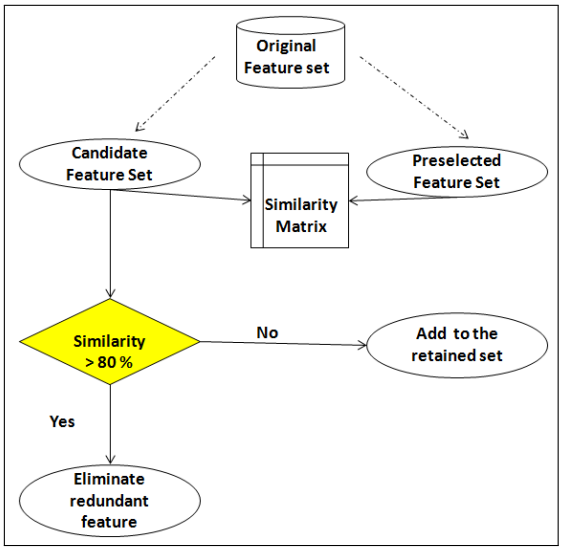

Fig. 1 First reduction process using prior knowledge.

3.2 Random Search Using Genetic Algorithms

Finding the best feature candidates from the reduced set can be seen as an enumeration problem. A random search algorithm refers to an algorithm that uses some kind of randomness or probability in the definition of the method. The term metaheuristic is also commonly associated with random search algorithms. Tabu search, evolutionary programming, ant colony optimization, GA [13, 14] and other random search methods are being widely applied to feature generation problems.

A GA use prior information to guide the search into the best region in the search space.

GA are better than conventional artificial intelligence (AI) algorithm in that it is more robust. GA is one of the artificial intelligence (AI) algorithms. However, unlike these older AI algorithms, a GA perform well even with noisy data or when the inputs changed slightly. Also, a GA may offer significant advantages over more usual search of optimization techniques especially in presence of large feature space.

GAs, are general adaptive optimization search methodologies that were developed by [15] to imitate the mechanism of genetic models of natural evolution and selection. They are a promising alternative to conventional random search methods. They work on the basis of a set of candidate solutions. Each candidate solution is called a “chromosome”, and the whole set of solutions is called a “population”. The algorithm allows movement from one population of chromosomes to a new popula-

tion in an iterative way, until acceptable results are obtained. Each iteration is called a "generation". A fitness function assesses the quality of a solution in the evaluation step. The crossover and mutation functions are the main operators that randomly impact the fitness value. Chromosomes are selected for reproduction by evaluating the fitness value. The fitter chromosomes have higher probability to be selected into the recombination pool using the roulette wheel or the tournament selection methods. Fig. 2 depicts the GA evolutionary process mentioned above.

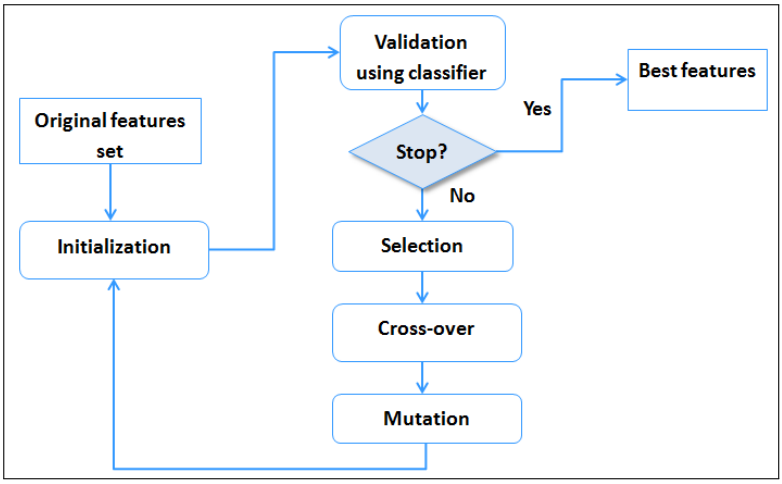

Fig. 2 General scheme for genetic algorithms.

4 Experimental Investigations

The experiments were conducted on Central Nervous System (CNS), a large data set concerned with the prediction of central nervous system embryonal tumor outcome based on gene expression. This data set includes 60 samples containing 39 medulloblastoma survivors and 21 treatment failures. These samples are described by 7129 genes [16]. We consider also the Leukemia microarry gene expression dataset that consists of 72 samples which are all acute leukemia patients, either acute lymphoblastic leukemia (47 ALL) or acute myelogenous leukemia (25 AML). The total number of genes to be tested is 7129 [3]. Table 1 displays the characteristics of the datasets that have been used for evaluation.

Our wrapper approach is used with GA as random search technique wrapped with two different classifiers namely support vector machine (SVM) and decision tree (DT).

For GA, population size is 100, number of generation is 10 as terminating condition, crossover rate is 0.7 and mutation rate is 0.001. In order to study the performance of the proposed approach, several evaluation measure derived from the confusion matrix were used [17]. these evaluation measures are: the percentage of correct positive predictions (Precision), the percentage of positive classified instances that were predicted

Table 1 Datasets summary

Names	CNS	Leukemia
Total instances	60	72
Total features	7129	7129
Number of classes	2	2
Missing Values	No	No

as positive (Recall). Our new approach is compared to forward and backward wrapper feature selection and results are summarized in Tables 2 and 3.

5 Results Analysis

The results in Tables 2 and 3 show that the relevant attributes identified by the various wrappers have indeed improved classification precision and recall of DT and SVM when compared to classification precision and recall with all the inputs. In fact using forward and backward feature selection as a wrapper improved significantly the classification performances compared to using all features, but most of the cases, experimental results show employing wrapper feature selection using GA and prior information enhanced the classification performances.

Table 2 Classification accuracy using wrapper feature selection approach for CNS dataset.

Wrapper approach	Number of Attributes	Precision (%)	Recall (%)
GA+ DT	396	91 (%)	76(%)
GA+ SVM	361	89(%)	82(%)
Forward feature selection	367	69(%)	72(%)
Backward feature selection	370	67(%)	75(%)
With all inputs	7129	49(%)	57(%)

Table 3 Classification accuracy using wrapper feature selection approach for Leukemia dataset.

Wrapper approach	Number of Attributes	Precision (%)	Recall (%)
GA+ DT	392	92 (%)	73(%)
GA+ SVM	373	86(%)	85(%)
Forward feature selection	360	62(%)	68(%)
Backward feature selection	358	63(%)	65(%)
With all inputs	7129	48(%)	59(%)

A closer look at Tables 2 and 3 shows that results are much better within the DT outputs. Actually, DT classifiers are sometimes considered as embedded methods. These kinds of methods essentially perform feature selection within the learning process, which means that they are able to select relevant features on their own: using their own search strategy and splitting mechanism. In other words DT classifiers select relevant features at two different stages. In the first stage features are selected by DT objective function individual and in the second features are selected by wrapper evaluation with GA. In this way, only features that are selected at both stages will form the final feature subset which is very likely to include features of high relevance.

6 Conclusion

In this work we propose a new approach for wrapper feature selection using genetic algorithm for random search and prior information. The motivation is to construct a more robust feature selection model with less complexity than usual search strategies. In a first part we investigated the effect of using prior information on the search space. Then we we conduct a random search on the reduced space of features using genetic algorithm. Results on two biological datasets show the performance of our approach.

References

1. Ben Brahim, A., Bouaguel, W., Limam, M.: 24. In: Combining Feature Selection and Data Classification Using Ensemble Approaches: Application to Cancer Diagnosis and Credit Scoring, pp. 517–532. Taylor & Francis (2014)
2. Schowe, B., Morik, K.: Fast-ensembles of minimum redundancy feature selection. In: Ensembles in Machine Learning Applications: Studies in Computational Intelligence, vol. 373
3. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**, 531–537 (1999)
4. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering **17**(4), 491–502 (2005)
5. Karegowda, A.G., Jayaram, M.A., Manjunath, A.: Article: Feature subset selection problem using wrapper approach in supervised learning. International Journal of Computer Applications **1**(7), 13–17 (2010). Published By Foundation of Computer Science
6. Chan, Y.H., Wing, W.Y.N., Daniel, S.Y., Chan, P.P.K.: Empirical comparison of forward and backward search strategies in L-GEM based feature selection with RBFNN. In: Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC), pp. 1524–1527 (2010)
7. Yun, C., Shin, D., Jo, H., Yang, J., Kim, S.: An experimental study on feature subset selection methods. In: Proceedings of the 7th IEEE International Conference on Computer and Information Technology. CIT 2007, Washington, DC, USA, pp. 77–82. IEEE Computer Society (2007)

8. Martínez, H.P., Yannakakis, G.N.: Genetic search feature selection for affective modeling: a case study on reported preferences. In: Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments. AFFINE 2010, New York, NY, USA, pp. 15–20. ACM (2010)
9. Feature subset selection using a genetic algorithm. In: Liu, H., Motoda, H. (eds.): Feature Extraction, Construction and Selection. The Springer International Series in Engineering and Computer Science, vol. 453
10. Bonev, B.: Feature Selection based on Information Theory. Ph.D. thesis, University of Alicante (2010)
11. Kumar, G., Kumar, K.: A novel evaluation function for feature selection based upon information theory. In: Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 395–399 (2011)
12. Al-Ani, A., Deriche, M.: An optimal feature selection technique using the concept of mutual information. In: Proceedings of the Sixth International Symposium on Signal Processing and its Applications, pp. 477–480 (2001)
13. Zhang, H., Sun, G.: Feature selection using tabu search method. **35**(3), 701–711 (2002)
14. Ramirez, R., Puiggros, M.: A genetic programming approach to feature selection and classification of instantaneous cognitive states. In: Giacobini, M. (ed.) Applications of Evolutionary Computing. Lecture Notes in Computer Science, vol. 4448, pp. 311–319. Springer, Heidelberg (2007)
15. Holland, J.H.: Adaptation in natural and artificial systems. MIT Press, Cambridge (1992)
16. Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y.H., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., Golub, T.R.: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature **415**(6870), 436–442 (2002)
17. Okun, O.: Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations (2011)