A survey of variable selection methods and multiclass learning in bio informatics

[1]Megha Purohit, [2] Shyamal Pandya

[1]M.E student, [2] Research consultant
[1,2]Computer engineering,
[1]Gujarat Technological University, Ahmedabad, India
[2]Majesty IT Services, Ahmedabad, India

Abstract- **Feature selection based data mining methods is one of the most important research directions in the fields of machine learning in recent years. This paper presents a review of assorted feature selection methods named filter, wrapper and embedded and multiclass classifiers like support vector machines (SVM), decision tree, averaged perceptron and neural network. Additionally it conveys an assessment of classifiers for breast cancer dataset.**

***Index Terms* - Feature selection, Bio informatics, Machine learning, and Multiclass classification.**

I. INTRODUCTION

Bio-informatics now having the advancement of databases, algorithms, computational and statistical techniques and its forbearance towards the formal and sensible problems is mended by the management and evaluation of biological facts. The fields of machine learning are prospering by the feature selection which is based on the data mining methods. Especially in recent years, along with the disturbance of many high dimension/small sample problems, such as, natural language processing, biological data, economic and financial, network, telecom and medical data analysis the study of feature selection once again is acquiring the attention.

Formally, data mining methods are divided into two categories. The first one is unsupervised methods, e.g. cluster methods, expectation-maximization problem; they pay attention on the association of independent variables, response variables are not present at the time of cluster procedure. The other is supervised methods, such as logistic regression, decision trees and support vector machine and so on [1, 7]. They are dedicated to club independent variables and response variables. However, there are so many supervised data mining methods that it is difficult to resolve which one is coagulate better with the bio-informatics data. Therefore, comparison study of data mining methods is usually carried out to select an efficient method to revoke the bio-informatics issues.

Multiclass classification techniques may be approximately divided into two types. First are the binary classification algorithms that can be evidently extended to deal with multiclass issues. The other one is the separation of multiclass issues into binary ones. There is probably no multiclass technique that outperforms the whole set. The selection of the technique must be made relying on the constraints like the desired degree of accuracy, the time availability for development and training. It also depends upon which types of issues are arising. But, selecting the pleasant one is a very tough task.

II. FEATURE SELECTION BASED DATA MINING METHODS

Here three kinds of feature selection methods are used: filter, wrapper and embedded to carry out a comparison study to evaluate the better method suitable for biological dataset.

1 Filter

Filter techniques select variables without considering its type. Filter method gives supremacy to the least fascinating variables. The other variables will be a part of the model classification used to classify or statistics prediction. These techniques are specifically powerful in computation time and robust to over fitting [2]. Filter methods have also been used as a pre-processing step for wrapper methods, permitting wrapper to be used on large troubles. Although, filter techniques have a tendency to pick redundant variables due to the fact that they do not keep in mind the relationships between variables. Consequently, they are especially used as a pre-process method.

2 Wrappers

Excessive dimensionality is a first rate trouble for bio informatics dataset. One technique found to address this hassle is wrapper-based selection method. Wrapper methods train a new model for every subset; they are very computationally in depth; however they commonly provide the fine appearing feature set for that specific form of model. The fundamental premise of wrapper feature selection is constructing a model that is using a potential feature subset and the usage of the performance of this model as a score for the benefit of that subset. While constructing a model, a number of alternatives must be made in the way to build and compare the model. While this model may be constructed using the entire training set and then has its overall performance evaluated in opposition to that equal training set, this would potentially result in over fitting [3]. Wrapper strategies evaluate subsets of variables which permit, unlike filter approaches to discover the possible interactions between variables [4].

3 Embedded

Recently, embedded methods have been proposed so that reduction process in classification of machine learning can be possible easily. They are trying to mix the benefits of each preceding strategies. The machine learning algorithms take benefits of their own variable selection algorithms. So, it needs to be realized that a great selection is the one which limits their exploitation [5]. Partially because of the higher computational complexity of wrapper and a lesser degree embedded approaches, these strategies have not received good deals as long as filter proposals [6]. One thought is that we can reduce the usage of the multivariate filter methods and try to get improvement and have to increase the usage of wrapper and embedded methods.

III. MULTICLASS CLASSIFIERS

1 Neural network

An Artificial Neural Network is a data processing concept which is inspired by the way of biological nervous systems, such as the brain, process information. The basic element of this paradigm is the novel structure of this system. It is a combination of large number of highly interconnected processing elements (neurones). They work with unity to solve specific problems. A Neural Network has been configured for a specific application, such as pattern recognition or data classification, through a learning process. Neural network, due to its remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

2 Averaged perceptron

In machine learning, the perceptron is an algorithm for supervised learning of binary classifiers. Here functions that can decide whether an input belongs to one class or another. It is one kind of linear classifier, i.e. a classification algorithm that makes its predictions based on a linear predictor function combining a set of weights with the feature vector. The algorithm allows for online learning, in that it processes elements in the training set one at a time. In the context of neural networks, a perceptron is an artificial neuron using the Heaviside step function as the activation function. The perceptron algorithm is also termed the single-layer perceptron, to distinguish it from a multilayer perceptron, which is a misnomer for a more complicated neural network. As a linear classifier, the single-layer perceptron is the simplest feed forward neural network.

3 Decision tree

Decision trees are a powerful classification technique. Two widely known algorithms for building decision trees are Classification and Regression Trees [8] and ID3/C4.5 [9]. The tree tries to infer a split of the training data based on the values of the available features to produce a good generalization. The split at each node is based on the feature that gives the maximum information gain. Each leaf node corresponds to a class label. A new example is classified by following a path from the root node to a leaf node, where at each node a test is performed on some feature of that example. The leaf node reached is considered the class label for that example. The algorithm can naturally handle binary or multiclass classification problems. The leaf nodes can refer to either of the K classes concerned.

4 Support vector machines

Support Vector Machines are among the most robust and successful classification algorithms [10, 11]. They are based upon the idea of maximizing the margin i.e. maximizing the minimum distance from the separating hyper plane to the nearest example. The basic SVM supports only binary classification, but extensions [12, 13, 14, 15] have been proposed to handle the multiclass classification case as well. In these extensions, additional parameters and constraints are added to the optimization problem to handle the separation of the different classes. SVM is a very popular classifier in statistics and machine learning. SVM has several nice properties: 1) its dual formulation is relatively easy to implement (through Quadratic Programming). 2) SVM is robust to the model specification, which makes it very popular in various real applications [16].

IV. EVALUATION & COMPARISION

Breast Cancer Classification

Breast cancer is one of the most common tumours affecting women. Breast cancer patients with the same stage of disease can have distinctly different treatment responses and overall outcome. Breast cancer classification has been based primarily on the tumour, but with serious limitations. Breast cancer classification has been difficult in past because it has historical and specific biological insights, rather than systematic and unbiased approaches for recognizing tumour subtypes. In the context of the dataset the number n of features equal to the number of tumour-size (ranging from 1.5 to 2 centimetres) and the number N of samples is the number of patients under examinations (about hundreds).

Comparative study of classifiers

Confusion matrix is a matrix representation of the classification results. Table 1 displays the comparative accuracy between four classifiers.

Table 1: Comparison of supervised learning classifiers

Classifiers	Overall Accuracy	Average Accuracy
Neural network	0.639024	0.927805
Averaged perceptron	0.634146	0.926829
Decision tree	0.658537	0.931707
Support vector machine	0.64878	0.929756

V.CONCLUSION

This paper provides an observation on feature selection and multiclass classification for bio-informatics records. A comparison of various modern day classification techniques on multiclass dataset is demonstrated. The issue lies within the truth that the records are of excessive dimensionality and that the sample length is small. The prediction accuracy is dramatically lower for the datasets with a big wide variety of classes.

While increasing the range of samples is a possible approach to the problem of accuracy degradation, it is crucial to develop algorithms which can be able to investigate successfully multiple-class expression data for those unique datasets. Regardless of the good overall performance, the wrapper strategies have constrained methods due to the high computational complexity worried. This is genuine especially while the wrapper methods are carried out to Support vector machines (SVM), a kingdom-of-art classifier that has discovered success in a selection of methods.

REFERENCES

[1] [Delen D, Walker G, Kadam A, 2005]. "Predicting breast cancer survivability: a comparison of three data mining methods", Artif IntellMed, **34**:113–27.

[2] [J. Hammon, November 2013]. "Optimisation combinatoire pour la sélection de variables en régression en grande dimension": Application en génétique animale.

[3] [RandallWald, Taghi M. Khoshgoftaar]. "Optimizing Wrapper-Based Feature Selection for Use on Bioinformatics Data", Amri Napolitano Florida Atlantic University.

[4] [T. M. Phuong, Z. Lin et R. B. Altman, 2005]. "Choosing SNPs using feature selection. Proceeding, IEEE Computational Systems Bioinformatics Conference, pages 301-309.

[5] [B. Duval, J.-K. Hao et J. C. Hernandez Hernandez, 2009]. "A memetic algorithm for gene selection and molecular classification of an cancer". In Proceedings of the 11th Annual conference on Genetic and evolutionary computation, GECCO '09, pages 201-208, New York, NY, USA.

[6] [Yvan Saeys, Iñaki Inza and Pedro Larrañaga]. "Review of feature selection techniques in bioinformatics".

[7] [Huihui Zhao, Jianxin Chen, Y.Liu, Qi Shi, Yi Yang, Chenglong Zheng, 2011]. "The use of feature selection based data mining methods in biomarkers identification of disease", Elsevier, Beijing University of Chinese Medicine, China.

[8] [L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, 1984] "Classification and Regression Trees." Chapman and Hall.

[9] [J. R. Quinlan, 1993] "Programs for Machine Learning." Morgan Kaufmann.

[10] [Corinna Cortes and Vladimir Vapnik, 1995] "Support-vector networks." Machine Learning, pages 273–297.

[11] [C. J. Burges, 1998] "A tutorial on support vector machines for pattern recognition." In Data Mining and Knowledge Discovery, pages 1–47.

[12] [J. Weston and C. Watkins, 1998] "Multi-class support vector machines." Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London.

[13] [Erin J. Bredensteiner and Kristin P. Bennett, January 1999] "Multicategory classification by support vector machines." Computational Optimization and Applications, 12:53–79.

[14] [Koby Crammer and Yoram Singer, 2001] "On the algorithmic implementation of multiclass kernel-based vector machines." Journal of Machine Learning Research, pages 265–292.

[15] [Yoon kyung Lee, Yi Lin, and Grace Wahba, March 2004] "Multi category support vector machines: Theory and application to the classification of microarray data and satellite radiance data." Journal of the American Statistical Association, 99(465):67–81.

[16] [Xingye Qiao, Lingsong Zhang, August 2015] "Flexible High-Dimensional Classification Machines and Their Asymptotic Properties". Journal of Machine Learning Research 16 (2015) 1547-1572.