An Evaluation of Feature Selection Methods for Multiclass Learning in Bio Informatics

Megha Purohit
Student of Masters in Computer Engineering
SAL Institute of Technology & Engineering Research
Ahmedabad, India.

Pooja Mehta
Asst Professor
SAL Institute of Technology & Engineering Research Ahmedabad, India

ABSTRACT
Traditional data mining techniques such as classification or clustering have demonstrated achievement in datasets which has multiple instances in singly relation but while extreme point of dimensionality or complex dependencies presents in the data it fails to offer accuracy and correctness. In solution to this, Feature (attribute/variable) selection techniques since last two decades have verified its requisites to improve speed, prediction and reduce computational cost of machine learners. In this paper review of assorted feature selection methods named filter, wrapper and embedded with each classifier like support vector machines (SVM), averaged perceptron and neural network is presented. Additionally it conveys an assessment of which FS approach works better for which classifier for breast cancer dataset.

Keywords
Machine Learning, Multi class classification, Feature Selection

1. INTRODUCTION
The field of machine learning is prospering by the feature selection which is based on the data mining methods. In recent years, many high dimension/small sample problems of areas such as, natural language processing, biological data, economic and financial, network, telecom and medical data analysis required to deploy feature selection before optimizing a supervised learning or unsupervised learning. There are several supervised data mining methods that it is difficult to resolve which one coagulates better with the bio-informatics data. Therefore, assessment of data mining methods is usually carried out to select an efficient method to revoke the bio-informatics issues. Correspondingly, there are many adaptations and versions of feature selection suggested by literature but it all depends on the data like finance, biological, astronomical etc. Therefore, evaluation of each approach is necessary to know which FS method can be used for particular classification. Numerous articles provided comparison either among classification methods or feature selection methods which can't confirm best combination of FS method and classifier. Furthermore, classification advancements like binary and multi class classifiers should be evaluated with feature selection method are hence, an analysis required that can better evaluate each classifier with each feature selection method.

2. FEATURE SELECTION AND CLASSIFICATION ADVANCEMENTS
Filter, wrapper and embedded methods are habitually used to carry out a comparison study to evaluate the better method suitable for biological dataset.

2.1 Filters
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.

2.2 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 commonly provide the fine appearing feature set for that appearing feature set for that specific form of model. The fundamental premise of wrapper feature selection is constructing a model that 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].

2.3 Embedded
Recently, embedded methods have been proposed to reduce the classification of machine learning. 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 realize that what a great selection is 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].

2.4 Classification
Consequently, Final best featured set is applied on either classification or clustering. Proposed exploration is focused on extremely admired and revolutionary supervised learning classification which is based on a model which can predict classes of instances from the data set. If we talk about medical data, supervised learning like decision trees, artificial neural networks, SVM (Support vector machine), regression tree, KNN (K Nearest Neighborhood) has proven fine results

[9, 10, 3]. A variety of classification techniques have been presented since past 25 years for medical applications. Classification methods were broadly classified into one class or binary classification, multi class classification and hierarchy multi class classification. Literature suggests that binary classification is barely credible for medical investigation like whether a patient has Rh +ve or –ve, viruses are present – Yes or No or else patient is male or female. Here classification property is resulting in to only two discrete values. Yet this will no longer help in the multi class problem such as a patients have many symptoms and each or many of it can belong to one or multiple diseases hence becomes even more challenging in micro array temporal data . To answer this, publications recommended Multi class classification which is contemporary and interesting for researchers. They were originally based on binary one means multi class classification procedure was optimized by breaking so many classes in to pair of twos. They are based on 1. Indirect approach which are one against one, one against one, all against all and directed acyclic graph SVM.
2. Direct approach attempt to find separate boundaries for all classes in one step [16, 17, 18]. Many articles came out based on these basic techniques for multi class classification [19, 20]. Even though they are being used widely have some downsides that they are capable to form only one measure at a time hence it consumes more computational power and even expensive. What's more is difficult and lengthy mathematical implementation [14]. 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 agreeable one is a very sturdy task.

3. NUMERICAL EXPERIMENTS

3.1 Data Set Details

In this experiment, we have two data sources:

- Breast Cancer Info
- Breast Cancer Features

The Breast Cancer Info data set contains some Meta data about the data set. Specifically, it contains 102,294 rows and 11 columns. We use the first 11 columns of this data set, including the following parameters as columns.

Table 1: Columns of Breast cancer Info dataset

No.	Column Name
1)	Label
2)	image-finding-id
3)	study-finding-id
4)	image-id
5)	patient-id
6)	left breast
7)	MLO
8)	x-location
9)	y-location
10)	x-nipple-location
11)	y-nipple-location

Basically, this data set contains the label and much ID information for each examination: image-finding-id, study-finding-id, image-id, and patient-id.

The Breast Cancer Features data set has 102,294 rows and 118 columns. It contains the features for each patient.
There is a one-to-one correspondence relationship between each row of two data sets. In our experiment, we use the label and ID information in Breast Cancer Info data set to split the Breast Cancer Features data set into training and test data sets.

3.2 Analysis and Results

In this section we present experimental results and confusion matrix for basic three multi class classifier- Neural Network, Decision Tree, Averaged Perception and SVM for above mentioned datasets. Before this, we have generated three different (Filter, wrapper and embedded) featured set of original data.
Now, Confusion matrix is a table used to show how fit a particular classification model is. The simulation tool we have used is Azure Machine Learning Studio (AMLS).

3.2.1 Neural Network

When we applied featured sets to Neural Network, wrapper gave most accurate multi class prediction.

Metrics

Overall accuracy	0.639024
Average accuracy	0.927805
Micro-averaged precision	0.639024
Macro-averaged precision	NaN
Micro-averaged recall	0.639024
Macro-averaged recall	0.270765

Fig 1 Experimental result for multi class Neural Networks

Matrix presented below states almost 64% of accuracy during experiment.

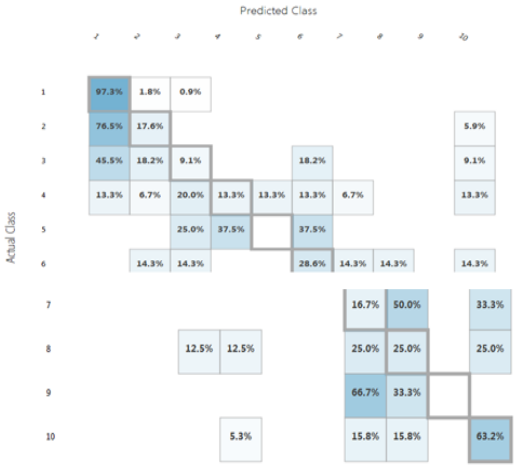

Fig 2 Confused matrixes for multi class neural network

3.2.2 Average Perceptron

Metrics

Overall accuracy	0.634146
Average accuracy	0.926829
Micro-averaged precision	0.634146
Macro-averaged precision	NaN
Micro-averaged recall	0.634146
Macro-averaged recall	0.233114

Fig 3 Experimental result for Averaged Perceptron

Multi class Average Perceptron offered 63% of accuracy and gave middling performance with all three types of FS methods. Wrapper performed better than other for this classifier.

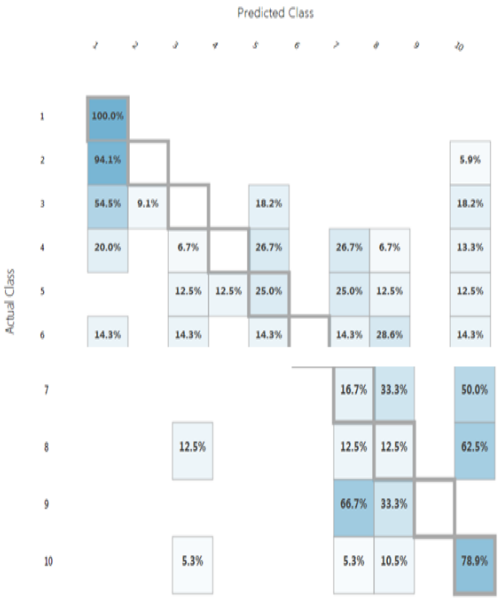

Fig 4 Confused matrixes for Multi Class Averaged perceptron

3.2.3 SVM

The third evaluation result is for the SVM. Featured set was inserted to SVM and it delivered overall 64.87% of precision.

Metrics

Overall accuracy	0.64878
Average accuracy	0.929756
Micro-averaged precision	0.64878
Macro-averaged precision	NaN
Micro-averaged recall	0.64878
Macro-averaged recall	0.273684

Fig 5 Experimental result for SVM

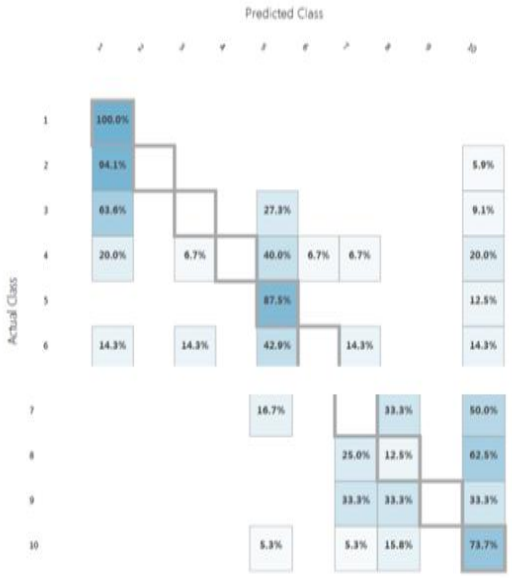

Fig 6 Confused matrixes for SVM

Table presented below narrates the generalized assessment report of FS method with classifiers for ALMS.

Essentially, three appraises are taken in to consideration accuracy, sensitivity and specificity for comparing each classifier for feature selection.

Table 2: Comparison of feature Selection methods meant for classifier

	NN	AP	SVM
Filter	Average	Average	Good
Wrapper	Good	Good	Very Good
Embedded	Average	Average	Average

Table 3: Performance statistics of Multi class prediction of Classifiers for featured set

Classifiers	Overall Accuracy	Average Accuracy
NN	0.639024	0.927805
AP	0.634146	0.926829
SVM	0.64878	0.929756

4. CONCLUSION

This paper provides observations on feature selection and multiclass classification for breast cancer data set records. A comparison of feature selection techniques with specific classification techniques on multiclass dataset is demonstrated. The major concern is that the records are excessive multi dimensional and sample size is small and the forecast accuracy is considerably inferior for the datasets with a good variety of classes. It is crucial to develop algorithms

which can be able to investigate multiple-class expression data for those unique datasets efficiently. Experiment results a wrapper with highest accuracy. Regardless of the overall performance, the wrapper strategies have constrained methods due to the high computational complexity. Wrappers acted very authentic with SVM multi class classifier. In addition if we formulate even minute change in sample a kingdom-of-art – SVM classifier that has performed better with a selection of methods.

5. 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]. "Optimization 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] Liu, Cutler G, Li W, Pan Z, and Peng S. "Multiclass cancer classification and biomarker discovery using GA-based algorithms." Bioinformatics 21(11) (June 2005): 2691-2697.

[9] Zhang, Min-Ling, José M. Peña, and Victor Robles. "Feature selection for multi-label naive Bayes classification." Information Sciences 179, no. 19 (2009): 3218–3229

[10] Dietterich, Thomas G., Richard H. Lathrop, and Tomás Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles." Artificial Intelligence 89, no. 1-2 (1997): 31–71

[11] Miller, A., B. Blott and T. Hames, 1992. Review of neural network applications in medical imaging and signal processing. Med. Biol. Engg. Comp., 30: 449-464

[12] Zaiane, Osmar R, Antonie Maria-luiza and A. Coman, 2001. Application of data mining techniques for medical image classification. Second Intl. Workshop on Multimedia Data Mining. In conjunction with ACM SIGKDD Conf. San Francisco, USA, Aug. 26

[13] Multiple Classifier Systems Lecture Notes in Computer Science Volume 3541, 2005, pp 278-285. Which Is the Best Multiclass SVM Method? An Empirical Study Kai-Bo Duan, S. Sathiya Keerthi

[14] F. Aiolli and A. Sperduti. An efficient SMO-like algorithm for multiclass SVM. In Proceedings of IEEE workshop on Neural Networks for Signal Processing, pages 297–306, 2002

[15] Weston, J. and Watkins, C., 1998, Multi-class Support Vector Machines. Royal Holloway, University of London, U. K., Technical Report CSD-TR-98-04

[16] Hsu, C.-W., and C.-J. Lin, C.-J., 2002, A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks, 13, 415-425.JAMES, G., 1998

[17] http://link.springer.com/article/10.1007/s10462-009-9114-9 A review on the combination of binary classifiers in multiclass problems

[18] X. Chen, X. Zeng, and D. van Alphen. Multi-class feature selection for texture classification. Pattern Recognition Letters, 27(14):1685{1691, 2006

[19] G. Madazrov and D. Gjorgjevikj. Evaluation of distance measures for multi-class classification in binary svm decision tree. In Artificial Intelligence and Soft Computing: 10th International Conference, (ICAISC), 2010.

[20] A.C.Lorena, A.C.Carvalho, J.M.Gama, are view on the combination of binary classifiers in multi-class problems, Artificial Intelligence Review30(1–4) (2008)19–37.

[21] L. J. van't Veer, H. Dai, and M. J. van de Vijver, Gene expression profiling predicts clinical outcome of breast cancer 2002. *Nature*, 415:530–536

[22] Dataset:http://www.rii.com/publications/2002/vantveer.htm