International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:3, No:10, 2009

Evaluation of the Impact of Dataset Characteristics for Classification Problems in Biological Applications

Kanthida Kusonmano, Michael Netzer, Bernhard Pfeifer, Christian Baumgartner, Klaus R. Liedl, and Armin Graber

***Abstract*—**Availability of high dimensional biological datasets such as from gene expression, proteomic, and metabolic experiments can be leveraged for the diagnosis and prognosis of diseases. Many classification methods in this area have been studied to predict disease states and separate between predefined classes such as patients with a special disease versus healthy controls. However, most of the existing research only focuses on a specific dataset. There is a lack of generic comparison between classifiers, which might provide a guideline for biologists or bioinformaticians to select the proper algorithm for new datasets. In this study, we compare the performance of popular classifiers, which are Support Vector Machine (SVM), Logistic Regression, k-Nearest Neighbor (k-NN), Naive Bayes, Decision Tree, and Random Forest based on mock datasets. We mimic common biological scenarios simulating various proportions of real discriminating biomarkers and different effect sizes thereof. The result shows that SVM performs quite stable and reaches a higher AUC compared to other methods. This may be explained due to the ability of SVM to minimize the probability of error. Moreover, Decision Tree with its good applicability for diagnosis and prognosis shows good performance in our experimental setup. Logistic Regression and Random Forest, however, strongly depend on the ratio of discriminators and perform better when having a higher number of discriminators.

***Keywords*—**Classification, High dimensional data, Machine learning

I. Introduction

HIGH-THROUGHPUT technologies have recently generated a large amount of data which enable analysis of a broad spectrum of biomolecules in a living cell [1]. For example, the transciptome, proteome, and metabolome can be studied by exploiting high-throughput datasets that comprise RNAs, proteins, and metabolites in a cell, respectively. Such sources of biological information facilitate the discovery, validation and commercialization of biomarkers. The main objective is to leverage these biomarkers for classification aiding the prognosis, prediction and diagnosis of a disease or treatment.

Supervised machine learning algorithms are very powerful methods and have been used for classification purposes [2]. In general, a model is built on significant patterns in training data and allows the prediction of states for future data. For example, microarray data can be used to construct a model for classifying disease states such as relating to cancer versus being healthy. To perform classification of biological data, many machine learning methods have been used so far on specific datasets [3, 4]. In addition, the performance of machine learning methods by using high-throughput data was also compared [5-7]. However, these studies are performed by using particular datasets. Hence, there is a lack of a comprehensive and systematic comparison between classifiers, which underscores the need of a guideline for biologists or bioinformaticians helping to determine an appropriate algorithm for a new dataset.

In this study, we use synthetic data to discover the impact of datasets characterized by their number of total features and proportion and strength of discriminators with the objective to systematically evaluate the performance of various classification algorithms.

II. Background

One major task in bioinformatics is the classification of biological datasets. The general process is to train classifier to recognize patterns from given labeled training samples and to classify novel samples with the trained classifier [7]. The essence in classification is to minimize the probability of error in using the trained classifier, which is referred to as the structural risk [8]. There are several popular machine learning methods used for classification:

K. Kusonmano is with the Institute for Bioinformatics, University for Health Sciences, Medical Informatics and Technology (UMIT), A-6060 Hall in Tirol, Austria & Faculty of Chemistry and Pharmacy, Leopold-Franzens-University Innsbruck, Innrain 52a, A-6020 Innsbruck, Austria (corresponding author to provide e-mail: kanthida.kusonmano@umit.at).

M. Netzer, B. Pfeifer and C. Baumgartner are with the Institute of Biomedical Engineering, University for Health Sciences, Medical Informatics and Technology (UMIT), A-6060 Hall in Tirol, Austria. (e-mail: michael.netzer@umit.at, bernhard.pfeifer@umit.at, christian.baumgartner@umit.at).

K. R. Liedl is with Theoretical Chemistry, Faculty of Chemistry and Pharmacy, Center for Molecular Biosciences, Leopold-Franzens-University Innsbruck, Innrain 52a, A-6020 Innsbruck, Austria. (e-mail: klaus.liedl@uibk.ac.at)

A. Graber is with the Institute for Bioinformatics, University for Health Sciences, Medical Informatics and Technology (UMIT), A-6060 Hall in Tirol, Austria. (e-mail: armin.graber@umit.at)

A. Support Vector Machines (SVM)

The idea behind the SVM algorithm can be explained based on four basic concepts: (i) the separating hyperplane, (ii) the maximum margin hyperplane, (ii) the soft margin, and (iv) the kernel function [9]. In principle, a SVM seeks a separating hyperplane in the data that produces the largest separation margin between two classes. The use of a kernel technique allows the linear separation in non linear classification problems.

B. Logistic Regression Analysis

Logistic Regression Analysis constructs a separating hyperplane between two classes of a dataset [10]. The class membership is predicted by a probability measure

$$P(\text{disorder} = 1) = \frac{1}{1+e^{-z}} \text{ and } z = b_0 + \sum_i b_i x_i ,$$

where $b_i x_i$ are the regression coefficients describing the size of the contribution of the risk factors and b_0 is the intercept, describing the value of z when the value of all risk factors is zero [11]. By default, the standard decision indicator is a cut-off value of P = 0.5 [12].

C. k-Nearest Neighbor (k-NN)

k-NN is a memory-based learning algorithm [2]. By giving a new query point *x*, k-NN finds *k* points in training instances, which are closest in distance to *x*. Euclidean distance or other distance measures could be used to measure similarity between *x* and points in the given training set. Then class *x* is labeled by majority voting of *k* nearest neighbors.

D. Naive Bayes

A Naive Bayes classifier is a probabilistic classifier based on Bayes' rule of conditional probability [13]. It is based on the assumption of class independence. Denote P(H) to be the probability of an event H. P(H|E) represents the probability of H conditional on an event E. Bayes' theorem is given by the equation

$$P(H|E) = P(E_1|H)\, P(E_2|H) \ldots P(E_n|H)\, P(H) / P(E),$$

where E_n is considered as a feature and H is a class.

E. Decision Trees

Based on training data, a Decision Tree is built as a binary classification tree [13]. Each internal node tests a feature to determine class which is labeled at leaf nodes. For new unlabeled instances, the prediction is made by a path from root to leaf node according to features properties of a new instance. The class of new instance is labeled when reaching the leaf node. To construct a tree, features in each node are selected from top to bottom by calculating the information gain of features, which reduces the entropy by separating instances.

F. Random Forest

The Random Forest is a machine learning method consisting of many decision trees and the output is based on the class of individual trees [14]. Each tree is trained by bagging data from a dataset. For classification prediction, each constructed tree in the forest is used for majority voting of output classes to determine the class of a new instance.

III. Method

A. Datasets

Small and a *high dimensional* synthetic dataset were generated. The small dimensional dataset comprises 100 features whereas the high dimensional dataset 1000. Both datasets are dichotomous with equal number of instances of 50 in the reference and comparison group. In a biological context, this means that the dataset contains 50 controls (control group) and 50 cases (treatment group). Formally, the dataset can be described as a set of tuples *T*, where $T = \{(c_j,m)|c_j \in C,\ m \in M\}$ with C = {comparison, reference}, where C is the set of class labels and M denotes the set of features (e.g. gene expressions or metabolite concentrations).

Similar to Hong *et al*. [15] the reference group (RG) can be defined as

$RG_f = \alpha + \varepsilon,$

and the comparison group (CG) can be described as

$CG_f = \alpha + \gamma + \varepsilon,$

where RG_f and CG_f is the value of feature f in the reference and comparison group, respectively. α is the mean value of the feature (standard deviation = 1) in the RG and CG, γ represents the relative effect size (shown in Fig. 1) between two groups and ε represents the normally distributed error. We define a *discriminator* as a feature with $\gamma \neq 0$ (γ = 2, 4, 6, 8, and 10). In biological data, the meaning of discriminator would be comparable to a biomarker. For the CG the number of discriminators n has to be greater than zero. In our experiment the percentage range of discriminator in the dataset was set from one to five.

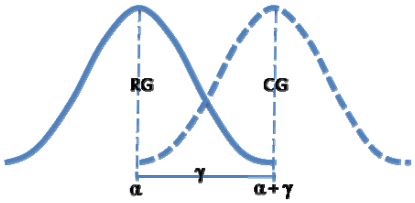

Fig. 1 γ denotes the effect size between two groups, RG and CG. With a given mean value of feature α in RG, the mean value of *discriminator* feature in CG is represented as α + γ without considering error ε.

B. Classification

We compared the discriminatory ability of six popular classifiers, which are SVM, Logistic Regression, k-NN, Naive Bayes, Decision Tree, and Random Forest based on simulated data in context of various proportions of discriminators and effect size. As an objective measure for estimating the discriminatory ability, we determined the area under the ROC curve (AUC) [16].

Weka [17] with mostly recommended default parameter settings listed in table 1 was used for exploring the performance of the previously outlined classification methods on our datasets. For the non-linear SVM, we used a polynomial kernel with an exponent of two.

In our experiments, we used a stratified 10-fold cross-validation strategy, where the dataset is subdivided into 10 roughly equal partitions and each in turn is used for validation and the remainder is used for training [2]. This process is repeated with 10 replications.

TABLE 1
LIST OF USED CLASSIFIERES AND PARAMETERS.

Name of classifier	Parameter setting
Decision Tree (C4.5)	Number of folds = 3
Random Forest	Number of trees = 9
Linear SVM	Complexity = 1, Epsilon = 1.0E-12, Kernel exponent = 1
Non-linear SVM	Complexity = 1, Epsilon = 1.0E-12, Kernel exponent = 2
Naive Bayes	Use Kernel Estimator = false
Logistic Regression	Ridge = 1.0E-8
K-Nearest Neighbor	Number of neighbors (K) = 5

IV. Results

The experiment was performed to investigate the performance of classifiers in context of various numbers of all features, various percentages of discriminators and effect sizes.

The performance of classifiers depicting various percentages of discriminators was investigated. By fixing value of effect size (γ=2), AUC of classifiers in both simulated small and high dimensional datasets are shown in Fig. 2-3, respectively.

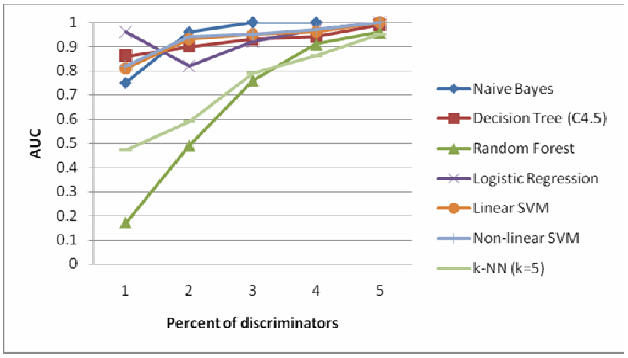

Fig. 2 AUC of classifiers in variation of increasing percentages of discriminators (features=100, γ=2).

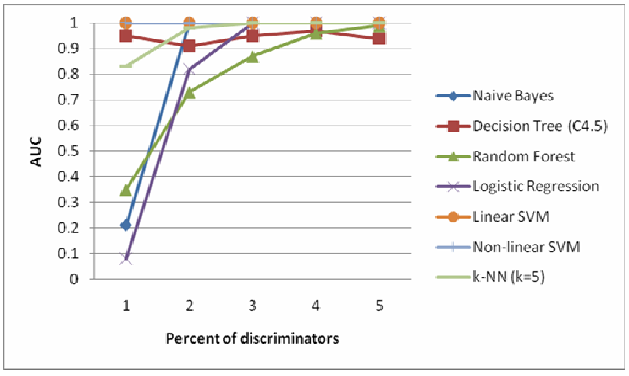

Fig. 3 AUC of classifiers in variation of percentages of discriminators (features=1000, γ=2).

In addition, the evaluation of classifiers in variation of effect size of discriminators (γ) was performed. In this case, the results of the two synthetic datasets with the proportion of discriminators fixed at 2 percent are shown in Fig. 4-5.

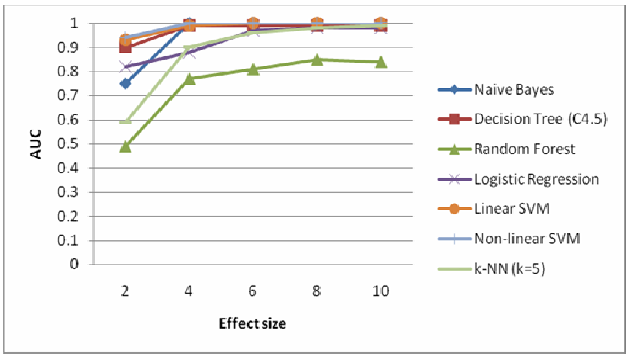

Fig. 4 AUC of classifiers in variation of effect size (features=100, number of discriminators=2%).

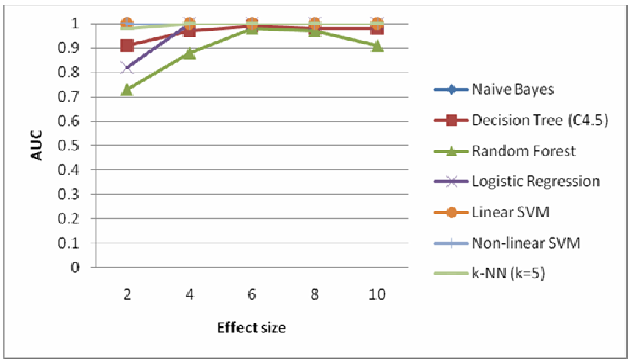

Fig. 5 AUC of classifiers in variation of effect size (features=1000, number of discriminators=2%).

Moreover, we also calculated the average AUC for each classifier by varying the effect size of discriminators for each percent of discriminators from 1 to 5% (Table 2-3). For example, for the small dimensional dataset (Table 2), the average AUC by varying effect size (γ = 2, 4, 6, 8, and 10) when percent of discriminators is 1 % using Non-linear SVM is 0.96. The overall average AUC of all proportions of discriminators (1-5%) is 0.99 with standard deviation (SD) of 0.02 as shown in the right most column of table. The results of higher percentages of discriminators (> 5%) are not shown since the different classifiers reach similar AUCs of approximately 1.

TABLE II
AVERAGE AUC OF CLASSIFIERS (FEATURE=100)

Algorithm/ Percent of discriminators	Avg. AUC 1	2	3	4	5	Avg± SD
Naive Bayes	0.84	0.95	**0.99**	**1.00**	**1.00**	0.96±0.07
Decision Tree (C4.5)	**0.96**	0.97	0.98	0.98	0.99	0.98±0.01
Random Forest	0.34	0.75	0.91	0.97	0.99	0.79±0.27
Logistic Regression	0.64	0.93	0.98	0.99	**1.00**	0.91±0.15
Linear SVM	0.95	0.98	**0.99**	0.99	**1.00**	0.98±0.02
Non-linear SVM	**0.96**	**0.99**	**0.99**	0.99	**1.00**	**0.99**±0.02
k-NN (k=5)	0.74	0.88	0.94	0.97	0.99	0.90±0.10

TABLE III
AVERAGE AUC OF CLASSIFIERS (FEATURE=1000).

Algorithm/ Percent of discriminators	Avg. AUC 1	2	3	4	5	Avg± SD
Naive Bayes	0.84	**1.00**	**1.00**	**1.00**	**1.00**	0.97±0.07
Decision Tree (C4.5)	0.98	0.97	0.98	0.96	0.98	0.97±0.01
Random Forest	0.43	0.89	0.97	0.99	**1.00**	0.86±0.24
Logistic Regression	0.12	0.96	**1.00**	**1.00**	**1.00**	0.82±0.39
Linear SVM	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**±0.00
Non-linear SVM	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**±0.00
k-NN (k=5)	0.96	**1.00**	**1.00**	**1.00**	**1.00**	0.99±0.02

As indicated in both Table 2-3, SVM gives the highest overall average AUC with low SD. Random Forest performance shows the lowest AUC and highest SD.

V. Discussion and Conclusion

In this work we compared the discriminatory ability of the classifiers SVM, Logistic Regression, k-NN, Naive Bayes, Decision Tree, and Random Forest based on simulated data representing various proportions and effect sizes of discriminators. The results show that SVM performs quite stable and reaches a high AUC compared to other methods. This may be explained by SVMs ability to minimize the structural risk when finding a unique hyper-plane with maximum margin to separate data from two classes. Yang et al. [8] described that this characteristic allows SVM the best generalization ability on unseen data compared with the other classifiers.

However, for diagnosis and prognosis of diseases, model-based classifiers such as Logistic Regression or classification trees are rather used than instance or kernel based methods [12], since the use of explicitly described equations and transparent rules is more practical and accepted for the daily clinical routine. For example, Decision Tree visualizes used features allowing the interpretation in biological context.

Nevertheless, Logistic Regression analysis did not perform well in our experimental setting when compared to other classifiers like SVM, which may be explained by the need of feature selection to extract only relevant features for building a model.

Decision Tree turned out to be very stable to different percentages of discriminators as well as diverse effect sizes. This can be explained due to the fact that Decision Tree relies on only features that lead to a low entropy and; furthermore, the entropy does not depend on the effect size since it does not affect purity.

Random Forest, however, strongly depends on the proportions of discriminators and performs better with a higher number of discriminators. This could be explained by the fact that the algorithm of Random Forest creates different subsets of features for training the tree. The higher the number of discriminators is, the higher the probability is that the tree is trained on discriminatory features.

The results of k-NN, Random Forest and Logistic Regression also show that feature selection approaches would be needed before classification since they demonstrate an inferior performance with a low number of discriminators.

Moreover, in our experimental setup we could not find a big difference in the performance of the classifiers between the small and the high dimensional dataset indicating no dependence on number of features.

The results also show that most differences of classifiers arise in the range of one to five percent of discriminators. Consequently, in biological application, there might be significant performance differences of classifiers on datasets with low number of biomarkers. The examination of this finding using biological data is part of our ongoing work. Furthermore, we plan to investigate the effect of feature selection approaches on classifier performances. Connections of feature selection approaches and classifiers will also be investigated.

Acknowledgment

K. K. thanks the Austrian Federal Ministry for Science and Research and the ASEA UNINET for providing the scholarship, and the University for Health Sciences, Medical Informatics and Technology (UMIT) for support during her study. This work was supported by the Austrian GEN-AU project Bioinformatics Integration Network. We also thank Matthias Dehmer from UMIT, who gave constructive comments to improve this manuscript.

References

[1] R. Clarke *et al.*, "The properties of high-dimensional data spaces: implication for exploring gene and protein and expression data", *Nature Reviews Cancer*, vol. 8, pp. 37-49, January, 2008.
[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Element of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2009.
[3] A. C. Tan, D. Q. Naiman, L. Xu, R. L. Winslow, and D. Geman, "Simple decision rules for classifying human cancers from gene expression profiles", *Bioinformatics*, vol. 21, pp. 3869-3904, August, 2005.
[4] R. Diaz-Uriarte, and S. Alvarez de Andres, "Gene selection and classification of microarray data using random forest", *BMC Bioinformatics*, vol. 7, January, 2006.
[5] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification", *BMC Bioinformatics*, vol. 9, July, 2008.
[6] M. Pirooznia, J. Y. Yang, M. Q. Yang, and Y. Deng, "A comparative study of different machine learning methods on microarray gene expression data", *BMC Genomics*, vol. 9, March, 2008.
[7] S. Cho, and H. Won, "Machine Learning in DNA Microarray Analysis for Cancer Classification", *Proc. of the First Asia-Pacific bioinformatics conference on Bioinformatics*, Australia, 2003, vol. 19, pp. 189-198.
[8] Z. R. Yang, "Biological applications of support vector machines", *BRIEF IN BIOIFORMATICS*, vol. 5, no. 4, pp. 328-338, December, 2004.
[9] M. Netzer, G. Millonig, M. Osl, B. Pfeifer, S. Praun, J. Villinger, W. Vogel, C. Baumgartner, "A new ensemble-based algorithm for identifying breath gas marker candidates in liver disease using ion molecule reaction mass spectrometry", *Bioinformatics*, vol. 25, pp. 941-947, April, 2009.
[10] D. W. Hosmer, and S. Lemeshow, *Applied logistic regression*, John Wiley and Sons, New York, USA, 2000.
[11] G. Tripepi, K. J. Jager, F. W. Dekker, and C. Zoccali, "Linear and logistic regression analysis", *Kidney International*, vol. 73, pp. 806–810, 2008.

[12] C. Baumgartner, and A. Graber, "Data mining and knowledge discovery in metabolomics", in F. Masseglia, P. Poncelet, M. Teisseire (eds.) Successes and new directions in data mining, Idea Group Inc., 2007, pp. 141-166.
[13] I. H. Witten, and E. Frank, *Data mining: practical machine learning tools and techniques*, Morgan Kaufmann, 2005.
[14] H. Pang, I. Kim, and H. Zhao, "Pathway-Based Methods for Analyzing Microarray Data", in F. Emmert-Streib, M. Dehmer (eds.) Analysis of Microarray Data, WILEY-VCH, 2008, pp. 356-358.
[15] F. Hong, and R. Breitling, "A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments", *BIOINFORMATICS*, vol. 24, no. 3, pp. 374–382, December, 2008.
[16] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado "The use of receiver operating characteristic curves in biomedical informatics", *Journal of Biomedical Informatics*, vol. 38, pp. 404-415, April, 2005.
[17] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka", *BIOINFORMATICS*, vol. 20, no. 15, pp. 2479-2481, April, 2004.