A Comparative Analysis of Feature Selection Methods for Biomarker Discovery in Study of Toxicant-treated Atlantic Cod (*Gadus morhua*) Liver

Xiaokang Zhang[0000−0003−4684−317X] and Inge Jonassen[0000−0003−4110−0748]

Computational Biology Unit, Department of Informatics, University of Bergen
{xiaokang.zhang, inge.jonassen}@uib.no
https://www.cbu.uib.no/jonassen/

Abstract. Univariate and multivariate feature selection methods can be used for biomarker discovery in analysis of toxicant exposure. Among the univariate methods, differential expression analysis (DEA) is often applied for its simplicity and interpretability. A characteristic of methods for DEA is that they treat genes individually, disregarding the correlation that exists between them. On the other hand, some multivariate feature selection methods are proposed for biomarker discovery. Provided with various biomarker discovery methods, how to choose the most suitable method for a specific dataset becomes a problem. In this paper, we present a framework for comparison of potential biomarker discovery methods: three methods that stem from different theories are compared by how stable they are and how well they can improve the classification accuracy. The three methods we have considered are: Significance Analysis of Microarrays (SAM) which identifies the differentially expressed genes; minimum Redundancy Maximum Relevance (mRMR) based on information theory; and Characteristic Direction (GeoDE) inspired by a graphical perspective. Tested on the gene expression data from two experiments exposing the cod fish to two different toxicants (MeHg and PCB 153), different methods stand out in different cases, so a decision upon the most suitable method should be made based on the dataset under study and the research interest.

Keywords: Feature selection · Stability · Classification · Biomarker discovery.

1 Introduction

Atlantic cod (*Gadus morhua*) is one of the most important commercial fish species in Norway [1], forming the basis for fisheries, trade, and, historically, civilization. Unfortunately, cod is increasingly susceptible to marine pollution from petroleum activities [2, 3]. Atlantic cod is commonly used as an indicator species in marine environmental monitoring programs, and a useful model organism to investigate the effect of toxicants [4–6]. Finding the best set of biomarkers

for Atlantic cod exposed to toxicants is of high research and commercial value. Biomarkers can for example be defined based on the expression level of a set of genes or proteins. Biomarker discovery is an essential part in study of toxicant exposure, and many methods have been proposed to find biomarkers [7]. However, a remaining question is, provided with numbers of biomarker discovery methods, which method is the most suitable one for a particular dataset. This paper provides a framework to compare potential biomarker discovery methods and to give researchers a better basis for choosing which one to use for the task at hand.

In the context of statistics and machine learning, biomarker discovery corresponds to a feature selection problem, where the purpose is to identify the most distinguishing features, for example, distinguishing normal and toxicant-treated cod livers. The task of feature selection is to identify, from a wide range of features, those that are best suited for classification.

The strategies of feature selection methods can be divided into two categories [7]:

1. Classical univariate statistical methods, where the features are considered as independent from each other. Genes that are differentially expressed are regarded as biomarkers.
2. Multivariate methods, which take the interaction between features into consideration when selecting the important features allowing to distinguish samples coming from different groups.

The classical univariate methods try to find the features having significantly different values between the different groups, e.g. control group and treated group. One of the most popular and basic methods is Student's t-test [8]. Some similar research also adopted Analysis of Variance (ANOVA) and Significance Analysis of Microarrays (SAM) to find the differential expressed genes [9–13]. A main drawback of such approaches is that they rest on the assumption that all the genes or proteins are independent from each other, which is clearly not true, since both genes and proteins are part of a biological system where they interact with each other [14, 15].

On the other hand, multivariate methods will take the interaction among features into consideration, reflecting that the features are acting in groups. Many feature selection and machine learning methods try to find the features most correlated with the class labels and take the interaction among features into consideration at the same time.

Feature selection methods are often divided into three categories: filter methods which focus on the relation between feature values and class labels; wrapper methods which use an objective function (can be the classification accuracy of the classifier) to evaluate features; and embedded methods where the classifier selects the features automatically [16]. The latter two are both classifier-dependent, and filter methods are more like a one-way decision without feedback from prediction accuracy. In order to find a more general feature selection method, which does not only work well with one specific classifier, we will only focus on the filter methods.

In toxicant exposure study, or more generally, in the context of biology, very often, researchers are faced with the high-dimension-small-sample-size issue, since it is hard and expensive to get a high number of samples (it is often around 10 or even lower), but the number of features (genes or proteins) is usually very high (over one thousand). In such cases, two problems are difficult to avoid: finding a reliable feature subset, as in this case the possibility of chance correlation is quite high; assuring that the selected features are true biomarkers. The true biomarkers should be data-independent, meaning that a small change in the samples should not lead to a large change in the selected features, which requires the feature selection method to be stable. Besides of that, they should also be qualified to be treated as the representatives of the whole feature list and should therefore be able to improve a classifier's prediction accuracy while classifying samples from different biological conditions. Therefore, we will compare the feature selection methods based on two aspects of their performance: stability to find a reliable feature subset and ability to improve a classifier's prediction accuracy.

To make the work reproducible, all the data sets and source codes are publicly available at https://github.com/zhxiaokang/FScompare.

2   Methods

2.1   Data sets

Two datasets from study of toxicant-treated Atlantic cod liver are used here. One is from the study of the hepatic proteome of MeHg-exposed Atlantic cod, where there are 10 samples in control group, 9 samples in low-dose treated group (0.5 mg/kg Body Weight MeHg), and 9 samples in high-dose treated group (2 mg/kg BW MeHg). The abundances of 1143 proteins were measured after the samples were exposed in vivo to MeHg for two weeks [12]. The other study is from the quantitative proteomics analysis of Atlantic cod livers treated with PCB 153 of various doses of PCB 153 (0, 0.5, 2 and 8 mg/kg BW PCB 153) for two weeks. There are 10 samples in each control group, low-dose treated group, medium-dose treated group, and high-dose treated group. Then 1272 liver proteins are quantified [13].

2.2   Principle of method and notations

Consider a set of m samples $\{x_i, y_i\}$ ($i = 1, 2, ...m$). Each sample has n input variables $x_{i,j}$ ($j = 1, 2, ...n$) and one output variable y_i. From the original feature set F, a feature selection method will select a subset S of k variables.

Suppose that there are P feature selection methods to be compared. Using Leave-One-Out Cross-Validation (LOOCV), m feature subsets will be generated for each pre-defined value of k. The stability of each feature selection method $Stab_{p,k}$ ($p = 1, 2, ...P$) can be calculated based on those m subsets.

To test their ability to improve a classifier's prediction accuracy, the generated feature subsets will then be applied to train a classifier and the prediction accuracy of the corresponding classifier will also be measured. Area Under the Curve (AUC) is used to measure the classifier's prediction accuracy [17]. If tested on Q classifiers, the prediction accuracy of each classifier can be calculated $AUC_{p,q,k}$ ($q = 1, 2, ...Q$). Considering both matrices $Stab$ and AUC, a general evaluation of each feature selection method can finally be achieved so that researchers can choose a proper method for their data.

But the stability does not necessarily agree with the prediction accuracy: the most stable feature selection method may not achieve the highest prediction accuracy. Then the researchers need to balance between these two measures according to their preference and the needs of the project.

2.3   Feature selection methods

Some representatives of those two strategies (univariate and multivariate) are compared. For the univariate methods, SAM is applied here, since it was used in the literature from where our data comes. SAM was designed to identify genes with significantly differential expression in microarray experiments. For the multivariate methods, we utilize minimum Redundancy Maximum Relevance (mRMR) [18] and Characteristic Direction from a geometrical aspect (GeoDE) [19]. mRMR is based on information theory. It tries to find out the feature subset in which the redundancy among the features are minimized and the relevance of features and the targeted classes are maximized. GeoDE uses linear discriminant analysis to define a separating hyperplane and the orientation of the hyperplane is used to identify the differentially expressed genes.

Those methods are selected for our comparison because they are based on different theories so that our results are more likely to be valid in general, and they are all widely used biomarker discovery methods. So P equals 3 in this case, but researchers can always compare as many feature selection methods as they want.

2.4   Performance measurement

Performance of feature selection methods is measured by two factors: stability and accuracy.

Many measures of stability have been proposed. Nogueira et al. studied 15 different measures proposed between 2002 and 2018 and also proposed their novel measure [20]. In our case where the purpose is to compare the stability of different feature selection methods, the absolute values of stability are not that important as long as they are comparable for different methods under the same settings. In each round of comparison, the number of selected features k is a constant, so the stability measure does not need to be able to cope with various numbers of features. LOOCV will generate more than two feature sets based on which the stability is calculated, so the measures which are defined for a pair of feature sets are not proper choices. Considering the measures that satisfy all the

requirements, we chose StabPerf [21] for its simplicity and interpretability. The stability is defined as:

$$Stab_{p,k} = \frac{\sum_{f \in F}(freq(f)/m)}{|F|} \tag{1}$$

Where $Stab_{p,k}$ is the stability of a given feature selection method p with a pre-defined k; m is the number of feature subsets analyzed; F is the set of features that appear in at least one of the m subsets and $|F|$ indicates the cardinality of F; $freq(f)$ is the frequency of feature $f \in F$ that appears in those m subsets.

To test the ability to improve a classifier's prediction accuracy, four popular classification methods are utilized here: Random Forest (RF) [22], Support Vector Machine (SVM) [23], and extended two-class logistic regression (RIDGE and LASSO are applied) [24].

2.5   Cross-validation approach

We characterize our problem as a two-class classification problem: the control group versus the treated group. In the process of classification, we need to divide the samples into training set and testing set. But since the number of samples is quite limited, we apply the strategy of LOOCV, which means that in every training-prediction process, we leave one sample out as testing set, and use the other samples as training set to search for the most important features and to train a classifier. With m samples, we will use the ith sample to test the prediction accuracy of the classifier trained from the other $m - 1$ samples. The average of performance observed over all m predictions will be regarded as the estimate of the performance of the model trained over the whole sample set. To avoid overfitting or an overly optimistic estimate, it should be noted that the feature selection and training of classifiers are only limited to the training set, to avoid the information from the testing set leaking into the model training procedure [25]. That makes the size of testing set decided by the number of samples in one classification problem, e.g. 19 in MeHg's high-dose case. Moreover, 19 samples indicate 19 rounds of feature selection and prediction, resulting in 19 selected feature subsets and 19*4 classifiers. Therefore, if a feature selection method is stable enough, there should be a big overlap among these 19 selected feature subsets; at best the feature subsets would be identical. And if the selected features are true biomarkers, the resulting 76 classifiers should yield high prediction accuracies.

To make our comparison more stable, avoiding the accidental findings, and to analyze the characteristic of the feature selection methods, we repeat the above process with different numbers of selected features (ranging from 40 to 400 with a step of 40, but also including 12 and 24 to look into more details with small numbers of selected features where the output varies a lot).

Tukey's Honestly Significant Difference Test (Tukey HSD Test) [26] is also applied to test the significance of the differences between different methods' performance on stability and prediction accuracy.
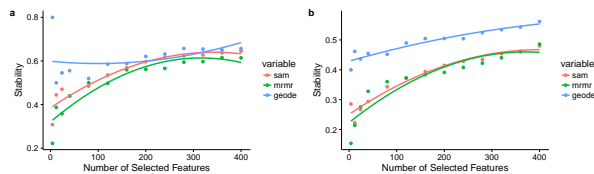
Fig. 1. Stability of feature selection methods on MeHg data. (a) Experiment on high-dose group versus control group. (b) Experiment on low-dose group versus control group.
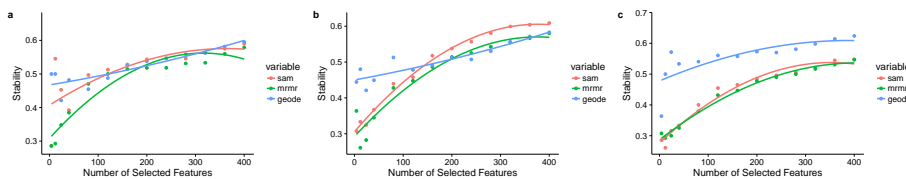
Fig. 2. Stability of feature selection methods on PCB 153 data. (a) Experiment on high-dose group versus control group. (b) Experiment on medium-dose group versus control group. (c) Experiment on low-dose group versus control group.

3   Results

3.1   Stability

We can see from Fig. 1 and Fig. 2 that the performance of GeoDE is more stable than SAM and mRMR across different numbers of selected features (with the smallest variance). Another big difference between GeoDE and the other two methods can be seen in low-dose condition of both MeHg and PCB 153: with all numbers of selected features, GeoDE consistently outperforms SAM and mRMR (Fig. 1b, Fig. 2c).

The results from Tukey HSD Test on stability are shown in Table 1. We limit the family error rate to 0.05, so the cases with an adjusted p-value (p-adj) smaller than 0.05 are regarded as significantly different. In accordance with the previous analysis, in low-dose condition both for MeHg and PCB 153, GeoDE is much more stable than the other two feature selection methods.

Table 1. Tukey HSD Test on stability

| Toxicant	Dose condition	Comparison	p-adj
MeHg	low	GeoDE is better than SAM	0.0006
MeHg	low	GeoDE is better than mRMR	0.0005
PCB 153	low	GeoDE is better than SAM	0.0014
PCB 153	low	GeoDE is better than mRMR	0.0007

Table 2. Tukey HSD Test on prediction accuracy

| Toxicant	Dose condition	Classifier	Comparison	p-adj
MeHg	high	RIDGE	mRMR is better than GeoDE	0.0107
MeHg	high	RIDGE	mRMR is better than SAM	0.0344
MeHg	high	LASSO	mRMR is better than GeoDE	0.0002
MeHg	high	RIDGE	SAM is better than GeoDE	0.0003
MeHg	low	LASSO	GeoDE is better than SAM	0.0004
PCB 153	high	LASSO	mRMR is better than GeoDE	0.0003
PCB 153	high	LASSO	SAM is better than GeoDE	0.0006
PCB 153	medium	SVM	mRMR is better than GeoDE	0.0077
PCB 153	medium	LASSO	SAM is better than GeoDE	0.0009
PCB 153	medium	LASSO	mRMR is better than GeoDE	0.0009
PCB 153	low	RF	GeoDE is better than mRMR	0.0002
PCB 153	low	RF	GeoDE is better than SAM	0.0082
PCB 153	low	SVM	GeoDE is better than SAM	0.0183

3.2 Accuracy

We find that the results of accuracy are not straightforward, since we will get different answers when asking which feature selection method performs the best. In each dose condition, all four classification methods are applied to assess the feature selection methods' ability to improve the prediction accuracy. Across different numbers of selected features, the AUCs of prediction are calculated. Fig. 3 is an example in the condition of low-dose MeHg. It shows that SAM performs the best when the classifier is SVM, but GeoDE turns out to be the best with the other three classifiers. To make it simple, for every experiment (each dose of each toxicant), we select the best classification method for it: a classifier that can give a high prediction accuracy for all three feature selection methods. For example, in low-dose condition of MeHg (Fig. 3), RIDGE gets the highest prediction accuracy compared with the other three classifiers regardless of the used feature selection method. Then Fig. 4 gives us all results for all conditions. As we can see, different feature selection methods stand out as the best. In low-dose condition of MeHg and PCB 153 (Fig. 4b, Fig. 4e), GeoDE performs the best, because it has a higher AUC than the other two in most cases of different numbers of selected features. For the other conditions, in high-dose condition of both MeHg and PCB 153 (Fig. 4a, Fig. 4c), and medium-dose condition of PCB 153 (Fig. 4d), mRMR stands out, especially with a low number of selected features.

Another phenomenon we can see from Fig. 4 is that based on gene expression data and our analysis, MeHg appears to influence cods more than PCB 153 does, since it is easier for classifiers to distinguish between control group and treated group with a small number of features (higher prediction accuracy), and the performance is also more stable.

According to the result of Tukey HSD Test on prediction accuracy (Table 2), in different dose conditions and with different classifiers, different feature selection methods will stand out. However, generally speaking, in high-dose condi-
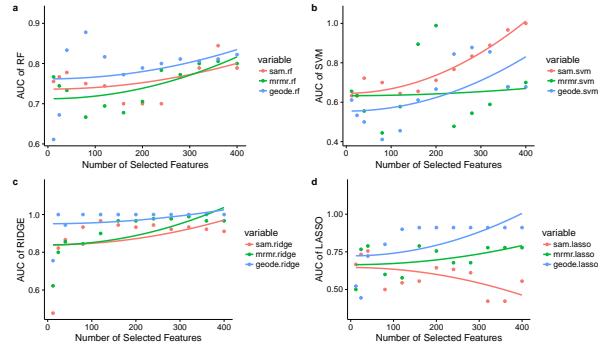
Fig. 3. Prediction accuracy on MeHg low dose data. (a) using RF (b) using SVM (c) using RIDGE (d) using LASSO.
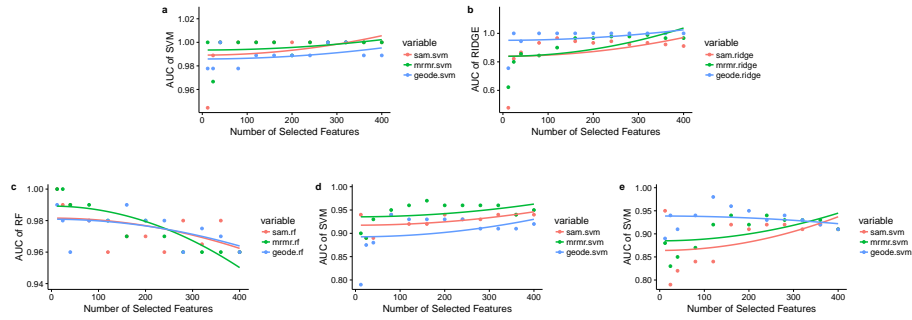
Fig. 4. Prediction accuracy. (a) in high-dose condition of MeHg (b) in low-dose condition of MeHg (c) in high-dose condition of PCB 153 (d) in medium-dose condition of PCB 153 (e) in low-dose condition of PCB 153.

tion, mRMR seems to outperform the other two feature selection methods, and in low-dose condition, GeoDE outperforms the other two.

4   Discussion and conclusion

In this article, we have presented a framework to choose the most suitable biomarker discovery method for a specific dataset by comparing the potential candidates from two aspects: stability, reflecting whether the selected feature subset is robust to changes in the training data, and resulting prediction accuracy.

On the aspect of stability to find a reliable feature subset, our results show that GeoDE is more stable than SAM and mRMR in two ways: its stability varies little across different numbers of selected features for all conditions, and the absolute values of stability are always the highest for all numbers of selected features in low-dose condition.

On the aspect of feature selection methods' ability to improve a classifier's prediction accuracy, in different dose conditions, different feature selection methods show up as the best. mRMR performs well in high-dose condition, but in low-dose condition, GeoDE outperforms the other two.

To conclude this case study, the choice of the most suitable biomarker discovery method quite depends on the dataset under study. If the experiments are conducted in high dose, then mRMR is the best choice, since it gives the highest prediction accuracy and its stability is comparable with the other two. If it's in low dose, then GeoDE is definitely the best choice, considering its excellent performance both in stability and prediction accuracy.

The framework of the comparative analysis is not limited to only this case study, but can be applied to any other similar study.

Acknowledgements

Funding

References

1. Ageeva, T.N. *et al.*: Gender-specific responses of mature Atlantic cod (Gadus morhua L.) to feed deprivation. Fisheries Research, 188, 95-99 (2017)

2. Goksøyr, A., Solberg, T. S., and Serigstad, B.: Immunochemical detection of cytochrome P450IA1 induction in cod larvae and juveniles exposed to a water soluble fraction of North Sea crude oil. Marine pollution bulletin, 22(3), 122-127 (1991)
3. Balk, L. *et al.*: Biomarkers in natural fish populations indicate adverse biological effects of offshore oil production. PLoS One, 6(5), e19735 (2011)
4. Sundt *et al.*: WCM 2010, 2012. NIVA, IMR, IRIS report (2012)
5. Chesman, B.S. *et al.*: Hepatic metallothionein and total oxyradical scavenging capacity in Atlantic cod Gadus morhua caged in open sea contamination gradients. Aquat Toxicol,84(3):310-20 (2007)
6. Olsvik, P.A. *et al.*: Are Atlantic Cod in Store Lungegrdsvann, a Seawater Recipient in Bergen, Affected by Environmental Contaminants? A qRT-PCR Survey. J Toxicol Environ Health Part A Curr Issues, 72(3-4):140-54 (2009)
7. Robotti, E., Manfredi, M., and Marengo, E.: Biomarkers discovery through multivariate statistical methods: a review of recently developed methods and applications in proteomics. J Proteomics Bioinform, 3, 20 (2014)
8. De Winter, J. C.: Using the Student's t-test with extremely small sample sizes. Practical Assessment, Research & Evaluation, 18(10) (2013)
9. Tusher, V.G., Tibshirani, R. and Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences, 98(9), pp.5116-5121 (2001)
10. Yadetie, F. *et al.*: Global transcriptome analysis of Atlantic cod (Gadus morhua) liver after in vivo methylmercury exposure suggests effects on energy metabolism pathways. Aquatic toxicology, 126, 314-325 (2013)
11. Yadetie, F. *et al.*: Liver transcriptome analysis of Atlantic cod (Gadus morhua) exposed to PCB 153 indicates effects on cell cycle regulation and lipid metabolism. BMC genomics, 15(1), 481 (2014)
12. Yadetie, F. *et al.*: Quantitative analyses of the hepatic proteome of methylmercury-exposed Atlantic cod (Gadus morhua) suggest oxidative stress-mediated effects on cellular energy metabolism. BMC genomics, 17(1), 554 (2016)
13. Yadetie, F. *et al.*: Quantitative proteomics analysis reveals perturbation of lipid metabolic pathways in the liver of Atlantic cod (Gadus morhua) treated with PCB 153. Aquatic Toxicology, 185, 19-28 (2017)
14. Shannon, P. *et al.*: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498-2504 (2003)
15. Tong, A.H.Y. *et al.*: Global mapping of the yeast genetic interaction network. science, 303(5659), 808-813 (2004)
16. He, Z., and Yu, W.: Stable feature selection for biomarker discovery. Computational biology and chemistry, 34(4), 215-225 (2010)
17. Fawcett, T.: An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874 (2006)
18. Peng, H., Long, F., and Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), 1226-1238 (2005)
19. Clark, N. R.*et al.*: The characteristic direction: a geometrical approach to identify differentially expressed genes. BMC bioinformatics, 15(1), 79 (2014)
20. Nogueira, S., Sechidis, K., and Brown, G.: On the Stability of Feature Selection Algorithms. Journal of Machine Learning Research, 18, 174-1 (2017)
21. Davis *et al.*: Reliable gene signatures for microarray classification: assessment of stability and performance. Bioinformatics, 22(19), 2356-2363 (2006)
22. Breiman, L.: Random forests. Machine learning, 45(1), 5-32 (2001)

23. Cortes, C., and Vapnik, V.: Support-vector networks. Machine learning, 20(3), 273-297 (1995)
24. Friedman, J., Hastie, T., and Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1), 1 (2010)
25. Cawley, G. C., and Talbot, N. L.: On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11(Jul), 2079-2107 (2010)
26. Yandell, B.: Practical data analysis for designed experiments. Routledge (2017)