Stable Feature Selection from Noisy Data

Dainius Jocas, Juozas Gordevičius

Vilnius University, Institute of Mathematics and Informatics

Abstract. Motivation:
Results:

1 Introduction and Motivation

Feature selection is an inevitable step when training classifiers for high-dimensional low-sample data. Such data is abundant in biomedical domain. For example, in gene expression experiments thousands of genes are interrogated using only a handful of samples per class. Other biomedical experiments, such as genome-wide association studies or epigenome analysis, yield millions of features whereas the number of samples remains uncomparably small. The advent of next-generation sequencing promises an explosion in the dimensionality of the data which will never be matched by the sample size. Therefore, future feature selection will remain an important topic in the foreseeable future.

Although many feature selection techniques have been proposed in the literature, they rarely (if ever) take into account the noise inherent in biomedical data. The data obtained through a complex chemical process is noisy. When the feature to sample ratio is large, noisiness leads to abundant random observations which will almost certainly skew the results.

The noise of the data leads to different feature selection results depending on the subsample of the data used. That is, feature selection is not stable with respect to data. Such instability can be seen in typical cross-validation results where classification accuracy may vary a little bit, but features used in each cross-validation fold will largely differ. We could show this result by example here. And we could argue that the issue is not so visible in other types of data where measurements are exact.

Stability of feature selection has so far received little attention. Loscalzo and Yu have proposed to discover dense feature groups and use one representative feature per group in further feature selection process. Other approaches instead concentrated on stability properties of different feature selection approaches assuming that stability is inherent in feature selection approach. We do not deny the assumption, instead we say that there is the other side of the meda.

In this paper we address the issues of classifying genetic data. The issues are:

- Stability of the selected features
- Time taken to train a classifier
- Quality of classification

Stability of features is important when we are interested in obtaining the minimal set of features that will lead to the best classification method. Typically, classifiers are evaluated using multiple folds of cross-validation. Within each fold, the features are selected using training dataset and the classifier is built. If each cross-validation fold yields a different set of features, then it is not clear which set of features should be used to build a final classifier.

The reasons why feature selection is not stable:

- The data is noisy and some features may seem informative due to random effects
- In high-dimensional low-sample data many features are likely to correlate. Which one of the correlating features will be selected by the feature selection method depends on the samples used in the current cross-validation run.
- Each feature selection method makes certain assumptions as to what makes a feature useful.

To summarize, the final set of features will depend on the training sample and the feature selection method. The more features there are in the input dataset, the more unstable will be the selection process.

Current approaches do not address all stability effects together. Fusion method allows to aggregate results from different feature selection methods thus alleviating the feature selection method bias. Consensus stable feature group method finds dense groups of features and uses only those features in the subsequent feature selection process.

We claim that these methods should be used together. Which looks trivially boring for the moment.

2 Related Work

In this section we review recent advances in classification literature giving special focus to classification in biomedical context. First we look at the classification methods: support vector machines, random forests, k nearest neighbours, etc. Second, we go through feature selection methods. Third, we look at the problem of feature stability.

When data is high dimensional and samples are few trained classifiers may easily overfit.

Classifiers learn the combination of features that separate best the sample groups. In case of biomedical gene expression data, the features correspond to genes and samples to individuals. Using the training data classifiers have to learn to distinguish between case and control individuals. The accuracy of the classifier is then evaluated by testing it on a separate test data. Also cross-validation is often used to evaluate the accuracy of a classifier [**?**]. Among the classical classification methods k nearest neighbours (kNN) [**?**] classifiers are widely used. Kernel approaches, such as Support vector machines (SVMs)[1] and kernel Fisher discriminant analysis are popular[2]. These methods, though, require the user to choose a kernel which is not a trivial task when no prior information about the data is available. Ensemble classification technique Random forests [3]

usually provides very good results without prior knowledge about the data.

Other popular solutions where only few predictive features have to be selected include logistic regression with ridge [4]. Other sophisticated penalty-based approaches are discussed by Zou and Li [5].

Gene expression data is typically analyzed using Support vector machines (SVM)[1], random forests [3], , or kernel Fisher discriminant analysis [2]. Theoretically they can be applied on any amount of dimencions yet in practice, the performance of these classifiers declines dramatically when the number of dimensions in the data exceeds significantly the number of samples. Furthermore, their runtime depends on dimensionality of the data and becomes practically not applicable in high-dimensional applications.

Feature selection will increase the accuracy of classifiers. Feature selection methods are of two types: filter and wrapper methods. In filter approach, first the features are selected using some external criteria and then the classifier is trained on them. These methods can be applied in tandem with any classification method. Wrapper methods are indistinguishable from the underlying classification method. For example SVM recursive feature elimination (SVM-RFE) uses the feature weights obtained from SVM classifier to remove the least significant features and recursively re-train the classifier on ever decreasing set of features [6]. Although wrapper methods often lead to better classification accuracy, they are also slower and may be difficult to parallelize.

Hua et al. have observed that in high-dimensional applications simple feature selection using Student's t-test will lead to classification accuracy comparable to that of SVM-RFE [7].

Tolosi et al. [8] observe that certain models that distribute weights to correlating features will not yield good classification results. If a group of such features is large enough, all the features ini it will appear irrelevant even if they yield high correlations with the outcome. Such methods are Lasso penalized logistic regression [9], group Lasso [10], fused SVM [11] and random forests [3]. The authors show that approaches where one feature per cluster is used avoid correlation bias.

Mention the review of stable feature selection for biomarker discovery

In stable feature selection review He and You observe that the method that should lead to the best result would use consensus clustering and consensus feature selection [12].

[13] define measures of robustness. One of the first works suggesting feature selection ensemble to improve robustness. The authors observe that robustness often comes at a price of classification accuracy. This is especially true for random forest classifiers.

A different approach to stability has been proposed by Yand and Mao [14]. They postulate that the instability problem arises from different assumptions about feature importance made by different feature selection methods. Therefore, by allowing the methods to vote they are able to obtain a consensus feature ranking that is quite stable. *** But not clear to me if they measure stability at all? Maybe they only measure performance of the resulting classifiers? ***. As the number of feature

selection methods is limited, the applicability of the approach is limited as well.
Feature selection on biomedical data often leads to unstable results. It has been observed, that different feature selection methods will select different features from the same dataset [14].
Park et al. group the features using a hierarchical clustering procedure and use the cluster centroids to train a linear regression method – sparse lasso [15].
Huang et al. propose a metagene method where features are clustered using k-means and principal components of the clusters are used to train the classification model [16, 17]. *** It is not clear to me how they make the final predictions. ***
Just as well, selecting features form different subsets of the same training dataset with the same feature selection method will likely yield different features [18]. In general, stability of feature selection is inversely proportional to dimenionality of the data. As the dimensionality increases, stability declines. Therefore, with the advent of next-generation sequencing and epigenome analyses that interrogate millions of loci, stability of feature selection remains an important issue.
Yu et al. have proposed to use kernel density estimation to discover dense groups of features [19]. Then, only the features close to the centers of the groups are used to train the classifier. This way, stability is increased and the number of features to select from is greatly reduced. However, the proposed mean shift algorithm takes $O(\lambda n^2 m)$ time where n is dimensionality of the data and m is the number of samples. (*** Wy care about samples? fishy. ***)
Consensus clustering introduced and discussed by Monti et al. [20].
Further, Loscalzo et al. proposed to combine dense group paradigm with consensus clustering [18]. They suggest to discover dense groups in multiple bootstrapped runs over the data and create a consensus set of groups for which the representative features will be selected. Although the approach increases feature stability, it is impossible to apply in practice due to overwhelming computational complexity.
Abeel et al. [21] propose a general framework for the analysis of robustness of a biomarker selection algorithm. They conduct large-scale analysis of ensemble feature selection techniques applying them on publicly available cancer gene expression datasets. As a measure of stability they use average of Kuncheva index (Kuncheva, 2007). As feature selection and classification method they use SVM-RFE with linear kernel.

3 Proposed Method

We propose to combine ensemble clustering and fusion of feature selection methods to obtain a fast, accurate and stable classifier.

1. Find the consensus clusters of given training data
2. Find features that represent the clusters
3. Use a feature selection method, to score the features
4. Test classification accuracy on test data

In our clustering step, we will find consensus clusters of correlating features. Then, for each such cluster, we will use one representative feature that correlates with the class label.

4 Exerimental Evaluation

Data. We use publically available gene expression datasets. First, we evaluate our methods on Cancer and CNS datasets that were also used to evaluate other works. Second, we use gene expression data from Stanley genomics website. The data investigates gene expression in the brain region called Brodman area 45. The samples include controls and schizophrenia as well as bipolar disease affected cases. Multiple datasets, interrogating same or different individuals are available. Thus, we are able to test our classification methods on independent data.

Overview of experiments. We perform the following experiments:

- Time complexity of different approaches. The aim is to show that our proposed approach is as fast as the fastest methods.
- Classification accuracy. The aim is to show using k-fold cross validation that our method yields classifiers that are more accurate than the other approaches. We further validate our findings on independent datasets.
- Stability. We show that features selected by our method are highly stable and more stable than those selected by other approaches.

4.1 Time Complexity

*** Dainiau, describe your experiment here ***

4.2 Classification accuracy

*** Dainiau, describe your experiment here ***

4.3 Stability of Feature Selection

*** We still have to do these experiments ***

5 Conclusions and Future Work

Yes, we rock! And in the future... we will rock even harder!

References

1. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3) (1995) 273–297
2. Cho, J., Lee, D., Park, J., Lee, I.: Gene selection and classification from microarray data using kernel machine. FEBS letters **571**(1-3) (2004) 93–98
3. Breiman, L.: Random forests. Machine learning **45**(1) (2001) 5–32
4. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning. Volume 1. Springer Series in Statistics (2001)

5. Zou, H., Li, R.: One-step sparse estimates in nonconcave penalized likelihood models. Annals of Statistics **36**(4) (2008) 1509
6. Li, F., Yang, Y.: Analysis of recursive gene selection approaches from microarray data. Bioinformatics **21**(19) (2005) 3741
7. Hua, J., Tembe, W., Dougherty, E.: Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognition **42**(3) (2009) 409–424
8. Toloşi, L., Lengauer, T.: Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics **27**(14) (2011) 1986–1994
9. Park, M., Hastie, T.: Penalized logistic regression for detecting gene interactions. Biostatistics **9**(1) (2008) 30–50
10. Meier, L., Van De Geer, S., Bühlmann, P.: The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(1) (2008) 53–71
11. Rapaport, F., Barillot, E., Vert, J.: Classification of arraycgh data using fused svm. Bioinformatics **24**(13) (2008) i375–i382
12. He, Z., Yu, W.: Stable feature selection for biomarker discovery. Computational Biology and Chemistry **34**(4) (2010) 215–225
13. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. Machine Learning and Knowledge Discovery in Databases (2008) 313–325
14. Yang, F., Mao, K.: Robust feature selection for microarray data based on multi-criterion fusion. Computational Biology and Bioinformatics, IEEE/ACM Transactions on (99) (2011) 1–1
15. Park, M., Hastie, T., Tibshirani, R.: Averaged gene expressions for regression. Biostatistics **8**(2) (2007) 212–227
16. Huang, E., Cheng, S., Dressman, H., Pittman, J., Tsou, M., Horng, C., Bild, A., Iversen, E., Liao, M., Chen, C., et al.: Gene expression predictors of breast cancer outcomes. The Lancet **361**(9369) (2003) 1590–1596
17. Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., Kloos, M., D'Amico, M., Pestell, R., West, M., Nevins, J., et al.: Gene expression phenotypic models that predict the activity of oncogenic pathways. Nature genetics **34**(2) (2003) 226–230
18. Loscalzo, S., Yu, L., Ding, C.: Consensus group stable feature selection. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2009) 567–576
19. Yu, L., Ding, C., Loscalzo, S.: Stable feature selection via dense feature groups. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2008) 803–811
20. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning **52**(1) (2003) 91–118
21. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., Saeys, Y.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics **26**(3) (2010) 392–398