Information Theoretic Feature Selection for High Dimensional Metagenomic Data

Gregory Ditzler and Gail Rosen
Dept. of Electrical and Computer Engineering
Drexel University
Philadelphia, PA 19104 USA
gregory.ditzler@gmail.com, gailr@ece.drexel.edu

Robi Polikar
Dept. of Electrical and Computer Engineering
Rowan University
Glassboro, NJ 08028 USA
polikar@rowan.edu

***Abstract*—Extremely high dimensional data sets are common in genomic classification scenarios, but they are particularly prevalent in metagenomic studies that represent samples as abundances of taxonomic units. Furthermore, the data dimensionality is typically much larger than the number of observations collected for each instance, a phenomenon known as curse of dimensionality, a particularly challenging problem for most machine learning algorithms. The biologists collecting and analyzing data need efficient methods to determine relationships between classes in a data set and the variables that are capable of differentiating between multiple groups in a study. The most common methods of metagenomic data analysis are those characterized by α– and β–diversity tests; however, neither of these tests allow scientists to identify the organisms that are most responsible for differentiating between different categories in a study. In this paper, we present an analysis of information theoretic feature selection methods for improving the classification accuracy with metagenomic data.**

I. Introduction

Next generation sequencers are providing researchers with copious amounts of data that need to be analyzed and classified into operational taxonomic units (OTUs), or some type of "class" associations such as a metagenomic phenotype. Automated classification tools can allow biologists rapid analysis of large microbial communities, and provide insights that may not have been possible without machine learning. In this work we focus on information theoretic feature selection, feature extraction and deriving relevant and discriminating features rather than heuristically tuning classifier parameters. Some applications of feature selection in genomics include:

Forensic Identification. Consider the scenario where we are trying to identify individuals based on the bacterial samples collected from their skin as presented by Fierer et al. application in microbial forensics [1]. Feature selection can provide biological information about what organisms are carrying the most mutual information between the organism and individual. Furthermore, there may exist powerful feature extraction methods that can also differentiate between individuals; however, the biological meaning is usually lost in the transformation of the features.

G. Ditzler and G. Rosen are supported by NSF CAREER #0845827, NSF Award #1120622, and DOE Award #SC004335. R. Polikar is supported by the NSF ECCS-0926159.

A Biological Study. Consider a data set collected from 1,000 healthy patients' and 1,000 unhealthy patients' guts. The goal is to determine which organisms carry information that can differentiate between the healthy and unhealthy populations. From a machine learning perspective, this is a feature selection problem; however, from a biological perspective, the selection of organisms allows the biologist the opportunity to examine why a set of species is responsible for differentiating healthy and unhealthy patients. It is important to note that there may be additional factors that influence the results, but may not be in the feature set.

II. Motivation

We focus on metagenomic data collected from short reads of 16S rRNA. Currently, biologists describe differences between groups in a metagenomic data set (e.g., controls and stimulus) using either α– or β–diversity metric in their biological studies. α–diversity measures variation within a class in the data set, whereas β–diversity measures variation between classes in a data set. β–diversity tools represent explicit comparisons of the microbial communities based on their composition[1]. Quantitative Insights Into Microbial Ecology (QIIME), which is a popular software tool for metagenomic data analysis that implements α– and β–diversity measures, uses distances between observations and principal coordinate analysis (PCoA) to implement β–diversity tools [2].

While QIIME offers many tools in aiding in the β–diversity analysis based on distances, it does not provide any tools that are based on finding information-theoretic methods to determine the most informative organisms in a data set that differentiate between the different groups in the data. In this section, we provide a motivation for information-theoretic feature selection to aid in: (*i*) improved risk bounds of a hypothesis (e.g., limiting the probability of error that can be made by the classifier) on data sets with large dimensionality, and (*ii*) find organisms that provide a (relatively) large amount of information for differentiating between classes in a metagenomic data set. Brown et al. provide an excellent overview and summary of findings from the last two decades of research in information-theoretic feature selection, which can be found

[1] http://qiime.org/tutorials/tutorial.html

in [3]. Rosen et al. present a comprehensive introduction of metagenomics from a signal processing perspective [4].

A. Where does the data come from?

Metagenomic sequences, or short reads, are a collection of 16S rRNA sequence fragments obtained directly from an environmental sample. Thousands of reads from 16S rRNA are typically collected from each observation and the sequences are cleaned by removing short and low quality score reads from the database. Then, sequences are clustered using using an algorithm such as CD-HIT [5], which forms a set of representative sequences for each cluster. Representative sequences are aligned and classified into OTUs. Thus, each data observation is comprised of thousands of reads from many organisms, where the same organism may have been detected multiple times. Hence, each data instance can be represented as a D-dimensional vector, denoted by $\mathbf{x}$, where D is the number of unique OTUs that were detected in a data set. Each element of $\mathbf{x}$ then represents the relative / absolute abundance of each OTU detected in the sample. For metagenomic data, D is typically very large (e.g., $D > 1,000$). One may view the random variable $\mathbf{x}$ as being distributed according to a multinomial distribution.

B. Problem Setting for Classification

The curse of dimensionality stems from the problem of learning a set of finite classes $\omega \in \Omega$ from a collection of observations $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^D$ when D is much larger than n. In designing a discriminative classifier, we seek to learn a function $h_\Omega : \mathbb{R}^D \mapsto \Omega$ such that $h_\Omega \in [0, 1]$ for binary classification problems. The generalization of h_Ω on unobserved data is quite important for any classification scenario; however, generalization with metagenomic data sets needs to be examined very carefully. As an example let us examine the generalization risk $R(h_\Omega)$ for a hypothesis, where h_Ω is derived from a class of linear functions. Let $h_\Omega \in \mathcal{H}$ be the hypothesis selected within the hypothesis class. We bound the risk of a hypothesis h_Ω that is an empirical risk minimizer (ERM). Then with probability $1 - \delta$ the upper bound on the risk is given by,

$$R(h_\Omega) \leq \hat{R}(h_\Omega) + \sqrt{\frac{32}{n}\left(d \log\left(\frac{2e \cdot n}{d}\right) + \log\left(\frac{4}{\delta}\right)\right)} \quad (1)$$

where $\hat{R}(h_\Omega)$ is the empirical risk from an n sample estimate, and d is the Vapnik–Chervonenkis (VC) dimension, which for a linear classifier is $d = D + 1$ [6], [7]. Fig. 1 shows the risk bound in Eq. (1) for a linear classifier that achieves $\hat{R}(h_\Omega) = 0.1$. This simple result leads to a profound realization of the primary problem faced in metagenemic classification: there needs to be a "sufficient" amount of data to assure the generalization of even the simplest linear classifier. A more complex classifier will require even more data to achieve the same bound because of a larger VC–dimension.

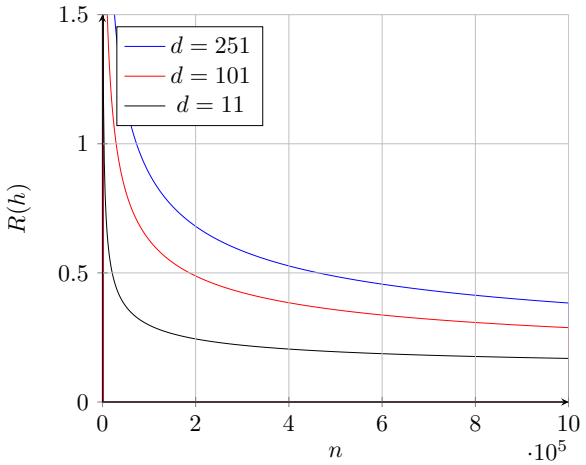

Fig. 1. Upper bound on the generalization risk of a hypothesis from a linear class of functions with $\hat{R}(h) = 0.1$ for problems in $\mathbb{R}^{250}$, $\mathbb{R}^{100}$, and $\mathbb{R}^{10}$.

TABLE I
FEATURE SELECTION METHODS USED FOR METAGENOMIC CLASSIFICATION

Criterion	Full Name
CIFE	Conditional Infomax Feature Extraction
CMIM	Conditional Mutual Info Maximization
JMI	Joint Mutual Information
MIFS	Mutual Information Feature Selection
MIM	Mutual Information Maximization
MRMR	Max-Relevance Min-Redundancy

C. What are we looking for in this study?

There are two primary questions that we seek to answer in this study as they relate to preprocessing, feature selection and extraction. First, is there a method – or perhaps a family of methods – that offers improvement to the accuracy of the classifiers? Clearly, expecting one method to perform better than all others is not feasible. Instead, we want to determine if information-theoretic methods choose features that improve classification accuracy, over heuristic feature selection methods, such as genetic algorithms. Second, is the choice of feature selection method robust to the selection of classification method?

III. METHODS

We have selected a set of data preprocessing and feature selection methods that are most commonly by the pattern recognition community, and conducted a rigorous comparative evaluation of them to determine their feasibility when applied to massive metagenomics data sets. In this section, we describe the tests performed and the data sets used in the study.

A. Feature Selection & Classification

In this work we test six information-theoretic feature selection methods, which can be found in Table I. Due to space constraints we do not delve into their theoretical implications or assumptions of the feature selection methods, the interested reader is encouraged to refer to literature on feature selection (e.g., see [3]). In conjunction with feature selection, we

TABLE II
RANK ON A FEATURE SELECTION METHOD FOR EACH DATA SET TESTED IN TABLE III. THE RANKS FOR EACH METHOD ARE THEN AVERAGED OVER THE 12 DATA SETS TESTED.

	brst	cln	leuk	lung	park	sonar	forns	biomeSite	biomeHost	costHost	costSite	costSex	Final
logistic regression													
none	2	4	4.5	5	1	1	4.5	7	7	7	7	6	4.67
cife	1	3	1	1	6	5	3	2	4	4	2	4	3.00
cmim	4	6.5	2	7	2	4	6	6	6	2	4	5	4.54
jmi	5.5	5	6.5	3	4	6	4.5	3	5	6	5	3	4.71
mifs	7	2	4.5	4	6	2	1	1	1	1	1	1	2.63
mim	5.5	1	3	2	3	7	2	4	2	5	3	2	3.29
mrmr	3	6.5	6.5	6	6	3	7	5	3	3	6	7	5.17
nearest neighbor													
none	7	1	1	4	4.5	4	4	7	7	7	7	7	5.04
cife	1	7	2	1	4.5	2	3	2	5	2	2	6	3.13
cmim	3	6	4	3	4.5	7	6.5	4	4	3	6	3	4.50
jmi	5	4.5	7	7	1.5	6	6.5	6	6	6	4	4.5	5.33
mifs	2	2.5	6	5	4.5	1	1	1	1	1	1	1	2.25
mim	4	2.5	4	2	1.5	3	5	3	3	4	5	2	3.25
mrmr	6	4.5	4	6	7	5	2	5	2	5	3	4.5	4.50
random forests													
none	7	2	1	4	4.5	6	4.5	7	7	7	5	6	5.08
cife	1	7	4	1	1	7	2	2	5	2	2	2	3.00
cmim	5	5	3	7	2	3	3	6	4	6	6	3	4.42
jmi	6	4	6	5	3	5	6.5	4	6	3	4	4	4.71
mifs	2	1	6	6	6	1	1	1	1	1	1	1	2.33
mim	4	3	2	2	4.5	2	4.5	3	3	5	3	7	3.58
mrmr	3	6	6	3	7	4	6.5	5	2	4	7	5	4.88
stochastic gradient descent													
none	5	1	1	7	6	7	5.5	7	7	7	7	6	5.54
cife	1	4	2	3	2	2.5	3	2	5	2	1	7	2.86
cmim	7	7	4	6	4	2.5	5.5	5	4	4	3	4	4.67
jmi	3.5	3	6.5	4	3	6	4	4	1	6	5	3	4.08
mifs	2	2	3	5	5	2.5	1	1	2	1	2	1	2.29
mim	3.5	5	5	2	1	5	7	6	3	3	4	5	4.13
mrmr	6	6	6.5	1	7	2.5	2	3	6	5	6	2	4.42

experiment with feature extraction as a precursor to feature selection. Feature extraction is form of dimensionally reduction; however, unlike feature selection, the resulting reduced feature set is determined via a mathematical transform. Thus, features in the reduced set with feature extraction will have a different physical meaning. We use Isomap, kernel PCoA, PCA and no feature extraction as the pre-processing algorithms. Furthermore, we select four classifiers – logistic regression, nearest neighbor, random forest and stochastic gradient descent – to test the consistency of the results.

Our implementation of the experiments are as follows: we first select a method for feature selection. Then for each classifier we run 10-fold cross-validation on each data set. We record the ranks of the classifiers based on accuracy, where low rank corresponds to high accuracy. The Friedman test is finally applied to the average ranks across all data sets for each feature selection methods. Refer to [8] for a detailed discussion on comparing different classifiers over multiple data sets (here we assume a "different classifier" is a "different feature selection method"). For brevity and consistency, we present the results in terms of the ranks rather than the numerical accuracy, because our statistical analysis is based on ranks and not the actual accuracy values of a classifier. Presenting the ranks is more consistent with our original goal to determine the feature selection methods that provide improvement over other feature selection methods listed in Table I.

Furthermore, in addition to the six feature selection methods, we also evaluated three different feature extraction methods: Isomap [9], kernel PCoA [10], [11], and PCA. The feature extraction methods are tested prior to feature selection.

B. Data Sets for Experimentation

There are a number of publicly available data sets for classification that are derived from metagenome studies. Most notably MG-RAST[2] allows users to download publicly released data sets. We use benchmark data sets for supervised classification of metagenomic data as presented by Knights et al. [12], which can be found on MG-RAST. The metagenomic data sets were download from MG-RAST with the species being the taxonomic level. We used the same metagenomic data sets as Knights et al. in their supervised classification work, and we add several data sets from the UCI machine learning repository [13]. Table III contains the data sets used for benchmarking purposes. Any data set indicated by a (†) is a metagenomic data set.

IV. EXPERIMENTAL RESULTS

In order to avoid clutter in the presentation of the results, we report the strictly the ranks of the classifier's classification

[2]http://metagenomics.anl.gov/

TABLE III
DATA SET INFORMATION. THE KEYS USED TO IDENTIFY THE DATA SETS ARE INDICATED BY PARENTHESIS.

Name	Cardinality	Features	Classes
breast (brst)	569	30	2
colon (cln)	62	2000	2
costelloHost (costHost)†	211	1858	9
costelloSex (costSex)†	211	1858	2
costelloSite (costSite)†	211	1858	6
forensic (forns)†	120	3695	3
leuk	72	7070	2
lung	73	325	7
microbiomeHost (bmeHost)†	1967	4123	2
microbiomeSite (bmeSite)†	1967	4123	4
parkinsons (park)	195	22	2
sonar	208	60	2

†Indicates a metagenomic data set.

TABLE IV
RANKS OF THE OVERALL ACCURACY OF THE FOUR CLASSIFIERS AVERAGED OVER SEVERAL DATA SETS DERIVED FROM TABLE III.

Processing	LR	NN	RFC	SGD	Final
Isomap	3.33 (4)	3.17 (4)	3.61 (4)	3.10 (4)	4.00
KPCoA	2.24 (2)	2.22 (2)	1.90 (2)	2.44 (3)	2.25
None	1.88 (1)	1.97 (1)	1.51 (1)	2.12 (1)	1.00
PCA	2.57 (3)	2.63 (3)	2.98 (3)	2.34 (2)	2.75

accuracy after 10 fold cross–validation as described in [8], where ranks range from 1 to 7 (i.e., the number of feature selection methods being tested). Fractional based ranking is used to deal with ties in multiple classifiers' accuracies.

Table II contains the ranks of the four selected classifiers on each data set under consideration. We present an overall rank for each feature selection method that is the average of a feature selection algorithm's ranks over all data sets (refer to the last column of Table II). The non-parametric Friedman test found that there are significant differences in the classification accuracies among the feature selection methods for each of the classifiers tested. One particularly interesting observation is that MIFS is working quite well on the metagenomic data sets. This result about MIFS is observed across several of the classifiers tested in this study and MIFS works particularly well on the metagenomic data sets (i.e., data derived from the forensic, Costello and the microbiome studies).

Finally, we also examine the impact of feature extraction as a pre-processing step to feature selection. We generate similar accuracy tables as Table IV for ISOMAP, KPCoA, and PCA. For each of the data sets and classifiers we rank the pre-processing methods rather than the feature selection methods (i.e., ranks are in [1,4]). We found that using the raw features in the feature selection routine consistently provided the best classification accuracy among the four methods tested. This result is not surprising as information is discarded by these feature extraction algorithms.

V. CONCLUSION

In this study we examined information theoretic feature selection for metagenomic data sets. We evaluated a number of information theoretic feature selection methods and data pre-processing methods on several benchmark metagenomic data sets. We found that using the raw features (i.e., no pre-processing) with feature selection improved the accuracy of the classifiers out of all the pre-processing methods tested. Furthermore, feature selection – all of which are derived from information-theoretic measures – improved the classification accuracy over no feature selection at all. This can be attributed to a large amount of irrelevant and redundant features in metagenomic data sets. Furthermore, MIFS appear to be quite robust for the metagenomic data sets, and all metagenomic data set used in this work is derived from taxonomic (i.e., species level) features.

Our current / future work is focused on feature selection for determining IBD patients based on metagenomic samples collected from the gut and consistency indices of the features selected. Furthermore, it would be worth while to investigate the ability of the feature selection methods remain consistent, as performed in [3], and examine the effect of the number of features selected (methods using redundancy should perform well when D gets large).

REFERENCES

[1] N. Fierer, C. L. Lauber, N. Zhou, D. McDonald, E. K. Costello, and R. Knight, "Forensic identification using skin bacterial communities," *Proceedings of the National Academy of Sciences*, vol. 107, no. 14, pp. 6477–6481, 2010.

[2] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight, "Qiime allows analysis of high-throughput community sequencing data," *Nature Methods*, vol. 7, pp. 335–336, 2010.

[3] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[4] G. L. Rosen, B. A. Sokhansanj, R. Polikar, M. A. Bruns, J. Russell, E. Garbarine, S. Essinger, and N. Yok, "Signal processing for metagenomics: Extracting information from the soup," *Current Genomics*, vol. 10, pp. 493–510, 2009.

[5] W. Li, L. Jaroszewski, and A. Godzik, "Clustering of highly homologous sequences to reduce the size of large protein databases," *Bioinfomatics*, vol. 17, no. 3, pp. 282–283, 2001.

[6] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[7] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verilag, 2nd ed., 1999.

[8] J. Demšar, "Statistical comparisions of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[9] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[10] B. Schlköpf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 1st ed., 2001.

[11] G. Ditzler, R. Polikar, and G. Rosen, "Forensic identification using environmental samples," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 1861–1864, 2012.

[12] D. Knights, E. K. Costello, and R. Knight, "Supervised classification of human microbiota," *FEMS Microbiology Reviews*, vol. 35, no. 2, pp. 343–359, 2011.

[13] A. Frank and A. Asuncion, "UCI machine learning repository." University of California, Irvine, School of Information and Computer Sciences, 2010.