BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 23 2011, pages 3242–3249
doi:10.1093/bioinformatics/btr547

Sequence analysis Advance Access publication October 7, 2011

Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data

Zhenqiu Liu[1,2,*], William Hsiao[2], Brandi L. Cantarel[2], Elliott Franco Drábek[2] and Claire Fraser-Liggett[2]

[1]Department of Epidemiology and Public Health, University of Maryland Greenebaum Cancer Center and [2]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Direct sequencing of microbes in human ecosystems (the human microbiome) has complemented single genome cultivation and sequencing to understand and explore the impact of commensal microbes on human health. As sequencing technologies improve and costs decline, the sophistication of data has outgrown available computational methods. While several existing machine learning methods have been adapted for analyzing microbiome data recently, there is not yet an efficient and dedicated algorithm available for multiclass classification of human microbiota.

Results: By combining instance-based and model-based learning, we propose a novel sparse distance-based learning method for simultaneous class prediction and feature (variable or taxa, which is used interchangeably) selection from multiple treatment populations on the basis of 16S rRNA sequence count data. Our proposed method simultaneously minimizes the intraclass distance and maximizes the interclass distance with many fewer estimated parameters than other methods. It is very efficient for problems with small sample sizes and unbalanced classes, which are common in metagenomic studies. We implemented this method in a MATLAB toolbox called *MetaDistance*. We also propose several approaches for data normalization and variance stabilization transformation in *MetaDistance*. We validate this method on several real and simulated 16S rRNA datasets to show that it outperforms existing methods for classifying metagenomic data. This article is the first to address simultaneous multifeature selection and class prediction with metagenomic count data.

Availability: The MATLAB toolbox is freely available online at http://metadistance.igs.umaryland.edu/.

Contact: zliu@umm.edu

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on April 4, 2011; revised on August 5, 2011; accepted on September 28, 2011

1 INTRODUCTION

The human body is inhabited by on the order of 10^{14} bacteria, collectively known as the human microbiota, which contains 100 times more genes (the microbiome) than in the human genome. Since these microbes interact with our bodies and provide functions lacking in our genome, changes in the microbial community structure are thought to impact our health. Given the vast number of genes in the microbiome and our inability to sequence all of them, a marker gene is often used for comparison among samples. The 16S rRNA gene is the most common marker gene used since it is universally present and well conserved among prokaryotes. Sequencing of 16S rRNA in an environment containing a mixed population allows the surveying of community structure without biases from culture-based methods at a relatively low cost. Whole genome shotgun (WGS) sequencing of the community (a metagenome), on the other hand, can provide estimates of functional capabilities of microbiome (Turnbaugh *et al.*, 2007), but the cost is substantially higher. A main promise of metagenomics is that it will accelerate the discovery of novel genes and new drug target and provide new insights into diseases with unknown etiologies (Qin *et al.*, 2010; Wooley *et al.*, 2010).

The first step of 16S rRNA metagenomic analysis usually involves the classification of sequences by organism to reduce the dimensionality of the dataset (from millions of sequences to thousands of organisms). In the process, the number of sequences classified to each organism is kept to provide an estimate on organism abundance. Classification of sequences are done by comparing sequences from a sample to 16S rRNA from known taxa or by clustering of similar sequences into operational taxonomic unit (OTU), which represent an unnamed taxon. The end result is a series of sequence read counts associated with taxa in the sample. A popular software package for assigning 16S rRNA to known taxa based on *k*-mer frequencies is called ribosomal database project (RDP) Classifier (Wang *et al.*, 2007). In addition to sequence classification, software such as Mothur (Schloss *et al.*, 2009) and QIIME (Caporaso *et al.*, 2010; Lozupone and Knights 2005) provides diversity metrics and sample comparison statistics allowing comparison of microbiome profiles using alpha (within-community) and beta (across-communities) diversities. For graphical comparison, MEGAN (Huson *et al.*, 2007) can compare the OTU composition of frequency-normalized samples (Mitra *et al.*, 2009; Huson *et al.*, 2009). From these analyses, we can figure out what organisms are in each of the samples or classes of samples. While these tools can be used to compare and cluster samples based on microbiome profiles, they do not allow for the identification of differentially abundant microbes in samples. Therefore, the important question of which organisms (by virtue of their presence/absence or relative abundance) distinguish one

*To whom correspondence should be addressed.

sample class from another cannot be answered by these software packages. This can be done using MetaStats (White *et al.*, 2009), but this software can neither identify multiple differentially abundant microbes simultaneously nor classify samples into multiple classes. MetaStats also suffers the multiplicity problem with multiple tests. Knights *et al.,* (2010) applied some existing machine learning approaches to the classification of microbiome data, but novel and efficient software is not yet available for multiclass classification of microbiome count data using supervised learning methods.

Machine learning for multiclass (and more general multilabel) classification has been applied to microarray analysis, text mining and image identification (Allwein *et al.*, 2001; Crammer and Singer 2001; Vens and Struyf 2008; Xu *et al.*, 2010; Liu *et al.*, 2010). The main objective of supervised learning is to predict the class of a future sample given the class and metagenomic count data. Most supervised learning methods fall into two general categories: instance-based and model-based learning. Instance-based learning (IBL), such as k-nearest neighbor (KNN) (Zhang and Zhou 2007), predicts the class of a sample with unknown class by considering the classes of k-nearest neighbors. It is more robust for data with unbalanced classes and is efficient for multiclass classification with a small number of features. However, accuracy diminishes with increasing irrelevant features because of the curse of dimensionality. On the other hand, model-based learning methods, such as support vector machine (SVM) and logistic regression, are mainly designed for binary classification. They are designed to separate two different classes as far as possible without considering the intraclass distances. Multiclass problems are often handled by combining binary classifier outputs, such as one class against the other (one versus one) or one class against the rest (one versus rest). However, when sample sizes are small, accuracy is reduced potentially due to noise and overfitting can occur since a high number of parameters needed to be estimated from a small number of samples [either $c(c-1)n/2$ or $(c-1)n$ parameters needed to be estimated with c classes and n features]. Furthermore, these methods also create unbalanced classification problems with the one versus rest rule even if the original dataset is balanced. Instance-based learning only takes into account the minimal distance, while model-based learning incorporates maximizing the interclass distances (e.g. maximizing the margin in SVM).

The integration of instance-based and model-based methods can maximize the interclass distances while minimizing the intraclass distances. Current integration (Cheng and Hullermeier 2009) only considers the labels of neighborhood instances as additional features for logistic regression, without utilizing the robustness of instance-based learning for unbalanced classes. Moreover, this method estimates many parameters and creates unbalanced classes in multi-class classifications, even if the original dataset is balanced. Because of the common issues associated with clinical samples: (i) small sample size and (ii) unbalanced classes, we propose a novel approach for multiclass classification through integrating instance-based and model-based learning to overcome these challenges in metagenomic data. Our proposed approach combines the KNN and SVM to simultaneously maximize the interclass distance and minimize the intraclass distance. This approach is robust for unbalanced classification, can classify multiple classes simultaneously without creating unbalanced classes and perform simultaneous feature (variable) selection and multiclass prediction with a simple parameter regulation, while estimating fewer parameters than previous approaches (only the same as the number of features). We apply our approach to 16S rRNA count data from metagenomic samples to select microbial taxa (features) that can distinguish one class of samples from others. The number of microbial taxa (features) is determined through cross-validation with smallest prediction error. We then use the selected features to build a weighted KNN classifier to predict a class for each sample. Because the dependence of the variance of the metagenomic count data for each taxon on the abundance of the taxa violates the homogeneity of variance assumption required for the application of many statistical methods, we developed variance stabilization methods to make non-homoskedastic count data easily tractable by standard machine learning methods. The current widely used data normalization method with proportion (relative abundance) only accounts for different levels of sampling across multiple individuals without adjusting for differences in variance. In this article, we describe several data normalization methods for variance stabilization before applying our proposed classification method. We evaluate the performance of our tool (*MetaDistance*) using simulated datasets and two publicly available real metagenomic datasets. The proposed methods are robust for all datasets and efficient for microbial feature identification and sample phenotype prediction.

2 METHODS

To understand the association between microbiota profiles and clinical phenotypes such as obesity, it is crucial to develop new supervised learning tools. We assume there are two or more populations with different clinical phenotypes (e.g. obese and lean, or different treatments and controls), each having multiple samples. We assume a set of non-overlapping taxa has been chosen, e.g. all genus-level groups appearing in the data. For each sample, we have one metagenomic count feature for each taxon, indicating the number of 16S rRNA sequence reads from the given sample assigned to that taxon, as shown in the following:

$$X=\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}, \text{ and } \qquad y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

where X is the metagenomic count matrix with n samples and m features, x_{ij} denotes the total number of reads assigned to feature j in sample i, and y is the clinical phenotypes with g categories. $y_i \in C=\{c_1,\dots,c_g\}$. Our goals are to identify features whose abundance in different populations is different, and to estimate the power of those identified features in predicting clinical phenotypes.

2.1 Data normalization and transformation

There are two sources of bias in the metagenomic count data: (i) different levels of reads (sampling) across multiple samples and (ii) dependence of the variance of x_{ij} on its particular value. The large the count value, the larger the variance. Validity of many statistical procedures relies upon the assumptions of normal distribution and homogeneity of variances. However, the metagenomic count and related percentage data have variances that are functions of the mean and are not normally distributed but instead are described by Poisson, binomial, negative binomial or other discrete distributions. The variance heterogeneity and non-normality of the metagenomic count data can seriously increase either Type I or II error and make the statistical inferences invalid (Kim and Taylor 1996; Kasuya 2004). Therefore, it is crucial to transform the count and percentage data prior to any standard analysis in order to correct deficiencies in normality and

homogeneity of variance (Freeman and Tukey 1950; Foi 2009; Olivier 2010). Our method for variance-stabilizing transformation and data normalization consists of two steps:

(1) Converting the raw abundance measure to a proportion (percentage) representing the relative contribution of each feature to each sample. This is to adjust for the sampling depth (read count) differences across samples. Mathematically, we normalize the metagenomic count matrix X into a proportion matrix P with

$$P=[p_{ij}]_{n\times m}, \qquad \text{where} \qquad p_{ij}=\frac{x_{ij}}{\sum_{j=1}^{m}x_{ij}}.$$

(2) We then employ either the square root transformation or the arcsine transformation to the metagenomic proportion matrix P (or original count matrix X):

- Square root transformation: This can be used either with the proportion matrix P or the original count matrix X, the transformed feature matrix $Z=[z_{ij}]_{n\times m}$ with

$$z_{ij}=\sqrt{p_{ij}+\frac{1}{2}} \qquad \text{or} \qquad z_{ij}=\sqrt{x_{ij}+\frac{1}{2}}$$

.

- Arcsine transformation: This is well suited for metagenomic proportion data P with

$$Z=[z_{ij}]_{n\times m} \qquad \text{with} \qquad z_{ij}=\arcsin(\sqrt{p_{ij}}).$$

This is very similar to the arcsine transformation with original count data X (Laubscher, 1961), which defined as

$$z_{ij}=\sqrt{L}\arcsin\sqrt{\frac{x_{ij}}{L}}+\sqrt{L-1}\arcsin\sqrt{\frac{x_{ij}+3/4}{L-3/2}},$$

where $L=\max(X)+4$ is the largest count value in count matrix X plus a constant number 4.

Before we do any transformations, we will compute the mean and variance for each sample with matrix P or X, and then test the assumption of homogeneity of variances with Bartlett's test (Nagarsenker, 1984). Either the square root or arcsine transformation will be used. Practically, if the percentage data have homogeneous variances, no transformation is needed. For data with variance heterogeneity, if the data lie in the range of 0–0.3 or 0.7–1 but not both, the square root transformation should be used. Otherwise, the arcsine transformation should be used. In most cases, we find both transformations increase predictive power and have similar performance. In this article, we therefore utilize the arcsine transformation with proportion data for all of our experiments.

2.2 Sparse-weighted distance learning with integrated KNN and SVM

A general multiclass classification problem may be simply described as follows. Given n samples, with normalized features, $D=\{(\mathbf{z}_1,y_1),\ldots,(\mathbf{z}_n,y_n)\}$, where $\mathbf{z}_i$ is a multidimensional feature vector with dimension m and g classes with class label $y_i\in C=\{c_1,\ldots,c_g\}$, find a classifier $f(\mathbf{z})$ such that for any normalized feature vector $\mathbf{z}$ with class label y, $f(\mathbf{z})$ predict class y correctly. Given two samples $\mathbf{z}_i$ and $\mathbf{z}_j$, we introduce a general weighted distance functions for KNN as follows:

$$\begin{aligned}D(\mathbf{w},\mathbf{z}_i,\mathbf{z}_j,p)&=w_1|z_{i1}-z_{j1}|^p+\cdots+w_m|z_{im}-z_{jm}|^p\\&=\sum_{r=1}^{m}w_r|z_{ir}-z_{jr}|^p=\mathbf{w}^t|\mathbf{z}_i-\mathbf{z}_j|^p,\end{aligned} \tag{1}$$

where $|.|$ denotes the absolute value, $w_k\geq 0$ for $k=1,\ldots,m$ are the non-negative weights and p is a positive free parameter. In particular, when $p=1$ and $p=2$, $D(\mathbf{w},\mathbf{z}_i,\mathbf{z}_j,1)$ and $D(\mathbf{w},\mathbf{z}_i,\mathbf{z}_j,2)$ represent the weighted city-block and Euclidean distances between $\mathbf{z}_i$ and $\mathbf{z}_j$, respectively. Given a new sample $\mathbf{z}_l$, we calculate KNN of $\mathbf{z}_l$ denoted by $N_k(\mathbf{z}_l,c_s)$ for each class c_s, and then take the average distance

$$D(\mathbf{z}_l,c_s)=\frac{\sum_{\mathbf{z}_i\in N_k(\mathbf{z}_l,c_s)}\{D(\hat{\mathbf{w}},\mathbf{z}_l,\mathbf{z}_i,p)\}}{k},$$

as the distance of $\mathbf{z}_l$ to class c_s. Finally, we assign $\mathbf{z}_l$ to a class c_j by means of a minimal distance vote.

$$\hat{y_l}=\arg\min_{c_j\in C}\{D(\mathbf{z}_l,c_j)\}.$$

2.3 Efficient quadratic SVM method for weight estimation

Now, the problem left is how to find optimal $\mathbf{w}$ for high-dimensional metagenomic data. As we discuss earlier, we want to choose $\mathbf{w}$ with small intraclass distance and large interclass distances simultaneously and automatically identify the features relevant to the phenotypes. We, therefore, propose an efficient quadratic SVM for weight estimations as follows:

$$\begin{aligned}&\min_{\mathbf{w},\xi}\sum_i\sum_j(\xi_{ij}^a)^2+\sum_i\sum_j(\xi_{ij}^b)^2+\lambda\sum_{k=1}^{m}w_k\\&\text{s.t. }\mathbf{w}^T(|\mathbf{z}_i-\mathbf{z}_j|_.^p)\leq 1+\xi_{ij}^a,\forall y_i,y_j\in c_s,\ \&\,\mathbf{z}_j\in N_k(\mathbf{z}_i,c_s)\\&\qquad\mathbf{w}^T(|\mathbf{z}_i-\mathbf{z}_j|_.^p)\geq 2-\xi_{ij}^b,\forall y_i\in c_s,y_j\in c_t,\ \&\,s\neq t\\&\qquad\xi_{ij}^a\geq 0,\ \xi_{ij}^b\geq 0,\ \text{and }\ w_k\geq 0,\forall i,j,k,\end{aligned} \tag{2}$$

where $|\mathbf{z}_i-\mathbf{z}_j|_.^p=[(z_{i1}-z_{j1})^p,\ldots,(z_{im}-z_{jm})^p]^T$ is an element-wise operation, and λ, k and p will be determined through cross-validation. In Equation (2), the first constraint represents the KNN intraclass distances, and we restrict them to a soft upper bound 1. The second constraint indicates the interclass distances with a soft lower bound 2. Hence, we can enforce a soft margin 1 between the intraclass and interclass distances. Therefore, the solution of Equation (2) will guarantee a small KNN intraclass distance and large interclass distance simultaneously. Finally, the reason we used KNN instead of all the samples in the same class for the first constraint is that samples in one class may have multimodal distributions. It is too stringent and unrealistic to require that all samples in one class have small distances. Equation (2) is equivalent to the following problem:

$$\begin{aligned}&\min_{\mathbf{w},\xi}E=\sum_{i\in c_s,j\in c_s}\sum I_a(\mathbf{w}^T|\mathbf{z}_i-\mathbf{z}_j|_.^p-1)^2+\ldots\\&\qquad+\sum_{i\in c_s,j\in c_t}\sum I_b(2-\mathbf{w}^T|\mathbf{z}_i-\mathbf{z}_j|_.^p)^2+\lambda\sum_{k=1}^{m}w_k\\&\text{s.t.}\qquad w_k\geq 0,\ \forall\, k=1,\ldots,m,\end{aligned} \tag{3}$$

where $I_a=\begin{cases}1 & \text{if } \mathbf{w}^T|\mathbf{z}_i-\mathbf{z}_j|_.^p>1\\ 0 & \text{otherwise.}\end{cases}$ $\forall\ \ y_i,y_j\in c_s$, and $I_b=\begin{cases}1 & \text{if } \mathbf{w}^T|\mathbf{z}_i-\mathbf{z}_j|_.^p<2\\ 0 & \text{otherwise.}\end{cases}$ $\forall\ \ y_i\in c_s$, and $y_j\in c_t$. Equation (3) is a much simpler truncated quadratic programming with non-negative constraints. It can be solved very efficiently, even if the problem has both large sample size and high dimension. The first-order derivative for Equation (3) is as follows:

$$\begin{aligned}\frac{\partial E}{\partial w_k}&=\sum_{i\in c_s}\sum_{j\in c_s}\sum I_a(\mathbf{w}^T|\mathbf{z}_i-\mathbf{z}_j|_.^p-1)|x_{ik}-x_{jk}|^p-\ldots\\&-\sum_{i\in c_s,j\in c_t}\sum I_b(2-\mathbf{w}^T|\mathbf{z}_i-\mathbf{z}_j|_.^p)|x_{ik}-x_{jk}|^p+\lambda\end{aligned} \tag{4}$$

Based on Equation (4) and $w_k\geq 0$, we implement a standard conjugate gradient method (Hager and Zhang, 2006) with non-negative constraints in *MetaDistance*. Because E is a convex optimization with a convex constraint,

a global optimal solution is guaranteed theoretically. The global minimum of E is reached when each element of w_k satisfies one of two conditions: either (i) $w_k > 0$ and $(\partial E/\partial w_k)|_{\hat{\mathbf{w}}} = 0$ or (ii), $w_k = 0$ and $(\partial E/\partial w_k)|_{\hat{\mathbf{w}}} \geq 0$. In the first case, the feature is identified as important by receiving a positive weight while the corresponding term in the gradient reaches zero. In the second case, the feature is eliminated as the corresponding term in the gradient remains positive even when w_k reaches zero, at the edge of the feasible region. Letting $g(\mathbf{w}) = \frac{\partial_{\mathbf{w}} E(\mathbf{w})}{\partial \mathbf{w}}$, we have the following iterative gradient algorithm for E maximization and optimal weight estimation:

Algorithm for optimal weight estimation: given p, k, λ and $\epsilon = 10^{-6}$, initializing $\mathbf{w}^1 = (\mathbf{w}_1^1, \mathbf{w}_2^1, \ldots, \mathbf{w}_m^1)^T$ with unweighted $\mathbf{w}_k^1 = 1$, for $k = 1, \ldots, m$. *Update* $\mathbf{w}$

- $\mathbf{w}^{t+1} = \mathbf{w}^t + \alpha^t d^t$, where t: the number of iterations and α^t: the step size.
- d^t is updated with the conjugate gradient method:

$$d^t = g(\mathbf{w}^t) + u^t d^t \quad \text{and} \quad u^t = \frac{[g(\mathbf{w}^t) - g(\mathbf{w}^{t-1})]^T g(\mathbf{w}^t)}{g(\mathbf{w}^{t-1})^T g(\mathbf{w}^{t-1})}.$$

Stop when $|\mathbf{w}^{t+1} - \mathbf{w}^t| < \epsilon$ or each element w_k satisfies the following two conditions: either (i) $w_k > 0$ and $(\partial E/\partial w_k)|_{\hat{\mathbf{w}}} = 0$ or (ii), $w_k = 0$ and $(\partial E/\partial w_k)|_{\hat{\mathbf{w}}} \geq 0$.

2.3.1 Evaluation criteria and choices of parameters The performance of *Metadistance* for multiclass classification is mainly evaluated with the prediction (test) error. The small the prediction error, the better the prediction accuracy. The average area under the ROC curve (AUC) is also used as a performance measure. AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The large the AUC, the better the model performance. AUC for each class versus the rest is calculated with the KNN distance between the test data and each class. Intuitively, if a test sample is from that training class, their distance will be small. The average prediction AUC is then used as a measure of overall model performance.

The free parameters λ, k and p are also determined by cross-validation with the smallest prediction error. The regular parameter λ controls the sparsity of the model. The larger the value of λ, the fewer microbial features will be selected. If λ is too small, there will be overfitting and little sparsity. If λ is too large, the produced classifier will be very sparse (very small number of features with non-zero weight) but have poor predictiveness. The optimal λ and the number of predictive features (variables) with non-zero weights are determined with smallest prediction error through 10-fold cross-validation. The parameters p and the number of nearest neighbors k are also decided by cross-validation. We limit $k = 1, 2, \ldots, \min n_i$, where min n_i is the smallest sample size for one class. Our computational experiments with simulation and real data show k is in the range 5–25 for the best performance. For simplicity, we choose $p = 1$ or 2 only in all the computational experiments, but other choices of p do improve the predictive power of our method. Users should feel free to choose different P values in their computations. Quadratic SVMs are implemented in the *MetaDistance* software.

3 RESULTS

Simulation Data: *in silico* metagenomic datasets were generated to contain five classes (groups) in four samples sizes (10, 20, 50 and unbalanced sample size with 10, 20, 30, 40 and 50 for each class, respectively). Datasets (x_{ij}) were generated from negative binomial (NB) distributions with different means and dispersion parameters. The means for NB are simulated from the Gamma distribution with a mean (μ) of 100 and variance (σ^2) of 1000. The variance of NB is $\sigma_{NB}^2 = \mu + \mu^2/\text{scale}$, where scale = 1. We simulated 1000 features for each sample from NB distributions, which contained the first 10 relevant features having different distributions with distinguished μs. We used 2-fold cross-validation to evaluate the method. First, we normalized the data with proportion and arcsin transformation, and then divided the data into training and test equal subsets. The training subset was used for model construction, while the test subset was used to evaluate performance. The model parameters k, p and λ are determined from only the training data with leave-one-out cross-validation. Each simulation was performed 100 times for each sample size (Table 1).

MetaDistance can identify differentiated features with high accuracy even if the sample size is small or unbalanced (Table 1). As the sample size increases so does frequency of correctly identified features. At 10 samples per class, MetaDistance identifies 80% of relevant features with over 72% accuracy and 30% of features with over 92% accuracy compared with 50% features with over 93% accuracy when there are 20 samples per class. Increasing the sample size to 50 leads to identification of 100% of relevant features over 93% accuracy. Our method performs well even if the dataset has highly unbalanced sample size. The method is able to identify 40% of relevant features accurately in all experiments with both sample size of 50 and unbalanced. The average number of features identified increases with the sample size for each class (Table 1). For example, a sample size of 50 for each class identifies 9.87 (of 10) relevant features on average. For comparison purpose, we

Table 1. Frequencies of correctly identified features with different sample sizes

Sample size/per-class parameters (λ^*, k^*, p^*)	10 (5, 6, 1)	20 (20, 5, 1)	50 (100, 20, 1)	Unbalanced (95, 5, 1)
w_1	72	60	100	100
w_2	73	93	100	99
w_3	31	18	93	91
w_4	38	10	93	98
w_5	92	98	100	100
w_6	94	68	97	94
w_7	76	94	97	100
w_8	72	96	99	99
w_9	99	99	100	99
w_{10}	75	98	100	100
Average no. of features selected	7.35	7.87	9.87	9.79

also apply F-test (ANOVA) to 100 simulation data with the sample size of 50 for each class. In all, 76–91 features are identified and the average number of features selected is 82.6 with $P < 0.05$. With $P = 0.00005$, 16–27 features are selected and the average number of selected features is 20.2, which indicate that the false positive rate is high even if we adjust the multiplicity problem for multiple comparisons with a very conservative Bonferroni rule. *Metadistance* identifies multiple features simultaneously without encountering the multiplicity problem and it is more accurate in identifying true predictive features than the statistical test.

The prediction error was calculated and compared with KNN and multinomial logistic regression (mlogit) in R (http://www.r-project.org/) (Fig. 1). *MetaDistance* outperforms KNN and mlogit in terms of prediction error rate. KNN had the highest error rate, likely due to the curse of dimensionality, making feature selection important when applied to high-dimensional data. Mlogit also performs more poorly than MetaDistance, due to unbalanced classification problem and small sample size. Since we observe a lower error rate in *MetaDistance* with larger sample sizes, these results suggest an increases of sensitivity and specificity as sample size increases. In addition, the average prediction error (0.22) with the unbalanced dataset is slightly better than the prediction error (0.23) with the dataset of sample size 50, indicating our method is robust with unbalanced data.

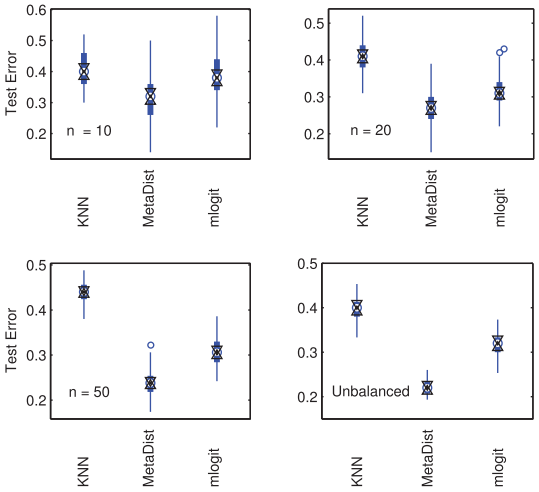

Fig. 1. Average prediction error with different sample sizes and different methods—Left: KNN; middle: MetaDistance; right: mlogit.

Benchmark metagenomic data: we applied our method to 815 16S rRNA metagenomic samples from six human body habitats (Costello *et al.*, 2009): external auditory canal (EAC), gut, hair, nostril, oral cavity (OC) and skin. Since the sample sizes range from 14 to 612 per habitat, this highly unbalanced dataset is dominated by one class (skin), which could create challenges for classification. Sequencing reads were classified into taxonomic groups using the RDP classifier using confidence threshold ≥ 0.5 1. About 5% of the sequences are not assigned to a genus by RDP. (56017/1070702 sequences are not assigned) (Wang *et al.*, 2007). The aim of the analysis was to identify taxonomic makers per habitat, whereas the original study used a binary classifier to determine whether samples originated from the gut, OC or other sites (Knights *et al.*, 2010). Using 100 permutations, we split the data into training (2/3 of samples) and test (1/3 of samples) and estimated parameters λ , p and k with 10-fold cross-validation with the training data only. Relevance count was calculated by the number of permutations a taxon is selected in a model. This analysis was performed at the bacterial family and genus levels of taxonomic assignment, with optimal parameters (λ^* , p^* , $k*$) of (410, 11, 1) for family and (450, 8, 1) for genus. The performances for this dataset are not sensitive to the parameter selections and the prediction errors are quite similar with a wide range of values for the parameters (Table 2).

MetaDistance identified 11 taxonomic markers at the family and genus level (Table 2), most of which have relevance counts of 100. The mean relative abundance (reads) of selected taxa across samples varies from 7 to 339 reads (column 3, 6). The abundance of these taxa could be used as markers to distinguish samples from different classes. Calculated prediction error rate of 0.075 (family) and 0.064 (genus) are smaller than reported for OTU analysis (Knights *et al.*, 2010). Using the abundances of these taxa, we are able to correctly classify 94.1% of samples to their correct body habitat (Table 3).

Table 2. Identified taxa and their relevance counts

Family level			Genus level		
Taxa	Relevance count	Mean reads	Genera	Relevance count	Mean reads
Chloroplast	100	58.3	*Neisseria*	100	27.5
Propionibacteriaceae	100	339.8	*Streptophyta*	100	58
Incertae Sedis XI	100	28.5	*Bacteroides*	100	35
Prevotellaceae	100	69	*Prevotella*	100	63.3
Corynebacteriaceae	100	80.9	*Staphylococcus*	100	126.2
Streptococcaceae	100	86.4	*Actinomyces*	100	33
Bacteroidaceae	100	35.2	*Alloiococcus*	88	23.7
Staphylococcaceae	100	131	*Streptococcus*	100	82.6
Pasteurellaceae	100	33.7	*Propionibacterium*	100	337.9
Carnobacteriaceae	100	34.5	*Corynebacterium*	100	65.5
Actinomycetaceae	100	33.5	*Peptoniphilus*	100	6.9
Test error	0.075 ± 0.017		Test error	0.064 ± 0.02	

This accuracy ranges by body habitat between 60% (Hair) to 100% (Gut). The average AUC is 0.99 across six classes. These results suggest that *MetaDistance* can accurately predict classes even with highly unbalanced sample sizes. For comparison purpose, we also analyzed the metagenomic OTU count data with similar procedure using only 552 non-transplanted samples (details in Supplementary Material). *MetaDistance* achieves the best predict error (0.08). We also compute the average prediction AUC based on the KNN distance between the test data and each class. The average prediction AUC is 0.99 with only 13 OTUs, compared with 27 OTUs reported by Knights *et al*. (2010). In addition, Gut and OC are perfectly separated from other classes, which is consistent with the result of Costello *et al*. (2009).

Metagenomic data of skin sites: using this same dataset, we repeated our analysis on 612 skin samples (Costello *et al.*, 2009), with the aim to classify each sample into a subhabitat (sample size)—Class 1: axilla (28), Class 2: external nose (14), Class 3: forehead (160), Class 4: glans penis (8), Class 5: labia minora (6), Class 6: lateral pinna (27), Class 7: palm (64), Class 8: palmar index finger (28), Class 9: plantar foot (64), Class 10: popliteal fossa (46), Class 11: umbilicus (12) and Class 12: volar forearm (155). This dataset represents several challenges: (i) a highly unbalanced classification ranging from 6 to 160 sample per subhabitat and (ii) previous methods have failed to separate compositional differences for these subhabitats. Compared with the one-versus-one strategy, which needs to estimate 66 models, only one model is needed with *MetaDistance* to identify features which are differentially abundant and capable of predicting classes. We divided the data into two parts: one with $\frac{2}{3}$ of the samples from each class as the training data and the remaining $\frac{1}{3}$ samples as the test data. The parameters λ, p and k were estimated using 10-fold cross-validation with the training data only. The parameter p has the choice of value 1 or 2 only, k is chosen from 1 to 20, and λ is selected from 1 to 40 with steps of 1. To prevent bias arising from a specific partition, we split the data 100 times and reported the relevance counts of the the identified taxa. The optimal parameters (λ^*, k^*, p^*) were (195, 3, 1) (family) and (210, 5, 1) (genus).

Table 3. Predicted cross-classification

		Predicted classes					
		EAC	Gut	Hair	Nostril	OC	Skin
True classes	EAC	11	0	0	0	0	3
	Gut	0	15	0	0	0	0
	Hair	0	0	3	0	0	2
	Nostril	0	0	0	12	0	3
	OC	0	0	0	0	14	4
	Skin	0	0	0	3	1	200

As shown in Table 4, we selected 12 taxonomic marker at the family and genus level assignment based on relative abundances, with test errors at 0.30 (family) and 0.31 (genus), comparable to previously reported best results (Knights *et al.*, 2010). The mean relative abundance (reads) of selected taxa across samples varies from 11 to 390 reads (columns 3, 6). Many of these markers are similar to those found in the habitat comparison, where skin samples were compared with other body habitats. Some of these taxa have been linked to disease, such as Prevotellaceae/Prevotella, which has been shown to be less prevalent in lean subjects (Zhang *et al.*, 2009). Additionally, species from the family Acinetobacter have been linked to disease and are target for health studies (Guner *et al.*, 2011).

We also plot a receiver operating characteristic (ROC) curve for each class versus the rest based on the KNN distance between the test data and each class from one run. Intuitively, if a test sample is from that training class, the KNN distance will be small. Otherwise, the distance will be large. Figure 2 shows the ROC curves and predictive AUC values at the bottom-right corner of each subplot. It is shown that Class 9: plantar foot is the easiest skin site to be separated from other classes with the prediction AUC of 0.99, and Class 2: external nose is the hardest skin site to be classified correctly with the test AUC of 0.74. The average test AUC for all classes is 0.88. The results indicate that the 12 identified taxa at family level have the predictive power for skin site discrimination. Obviously, it is more crucial to select important taxa that are highly discriminative for

Table 4. Identified taxa and their relevance counts

Family level			Genus level		
Taxa	Relevance count	Mean reads	Genera	Relevance count	Mean reads
Chloroplast	100	67	*Acinetobacter*	53	11.8
Propionibacteriaceae	100	392.5	*Veillonella*	57	40.5
Incertae Sedis XI	55	28.6	*Neisseria*	100	21.3
Prevotellaceae	80	58	*Streptophyta*	100	66.6
Corynebacteriaceae	100	74.6	*Prevotella*	100	53.5
Micrococcaceae	100	31.2	*Rothia*	91	18
Streptococcaceae	100	83	*Staphylococcus*	100	128.4
Veillonellaceae	85	47	*Pseudomonas*	51	11.6
Comamonadaceae	63	11.4	*Streptococcus*	100	78.3
Staphylococcaceae	100	133	*Propionibacterium*	100	390.1
Moraxellaceae	69	29.6	*Corynebacterium*	100	73
Neisseriaceae	97	22.4	*Fusobacterium*	77	13.2
Test error	0.30 ± 0.04		Test error	0.31 ± 0.03	

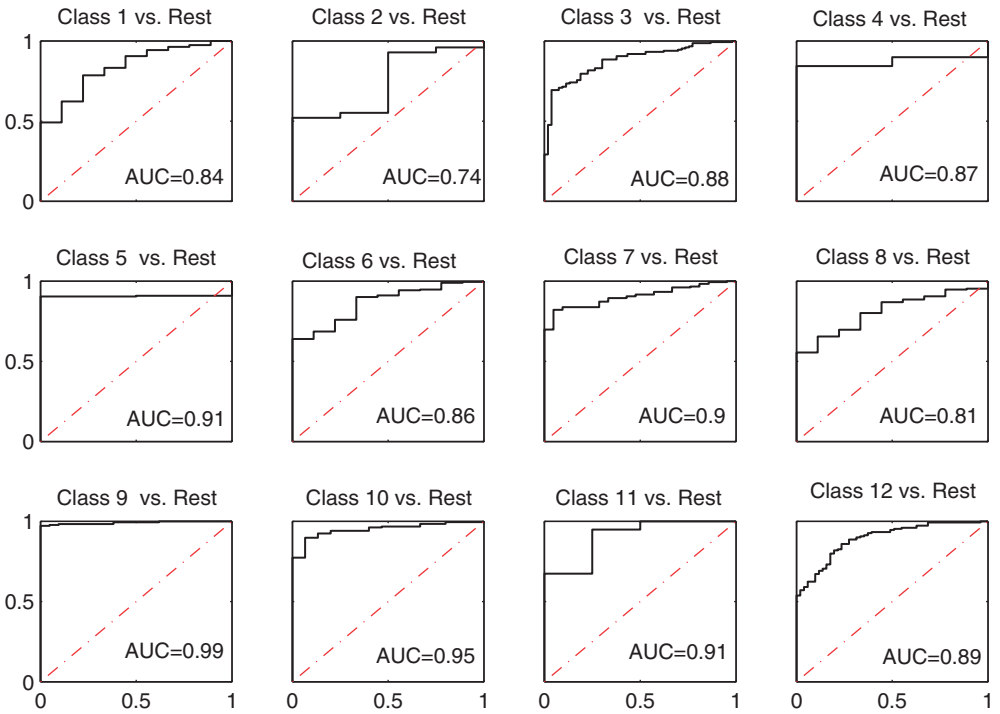

Fig. 2. ROC curves predicted with the eight selected taxa at family level.

this type of task, since the sites of sampling would most likely to be known. However, MetaDistance can certainly be applied for both taxa selection and class prediction with other types of metagenomic data where the category labels (such as disease status) are more expensive to obtain.

4 CONCLUSIONS

We have proposed a sparse distance learning method (MetaDistance) for multiclass classification through combining instance-based (KNN) and model-based (SVM) learning methods. The proposed method can identify phenotype-associated taxa and perform class prediction simultaneously. It is robust for unbalanced classification and can classify multiple classes simultaneously without creating unbalanced classes. In addition, this method estimates a small number of parameters (only the same as the number of features) and is very efficient for problems with small sample sizes, high dimensions and unbalanced classifications with many classes, which is common in genomic data. Experiments with limited simulation and real datasets demonstrated its effectiveness. While this method was tested on 16S rRNA, it can easily be applied to identify marker genes from WGS metagenomic and digital gene expression survey (SAGE) analysis without modification.

Funding: National Institutes of Health (1UH2DK083982-01 and 4UH3DK083991-02); National Cancer Institute (1R03CA133899-01A210).

Conflict of Interest: none declared.

REFERENCES

Allwein,E.L. *et al*. (2001) Reducing multiclass to binary: a unifying approach for margin classifiers. *J. Mach. Learn. Res.*, **9**, 113–141.

Caporaso,J.G. *et al*. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Crammer,K. and Singer,Y. (2001) On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, **2**, 265–292.

Cheng,W. and Hullermeier,E. (2009) Combining instance-based learning and logistic regression for multilabel classification. *Mach. Learn.*, **76**, 211–225.

Costello,E.K. *et al*. (2009) Bacterial community variation in human body habitats across space and time. *Science*. **326**, 1694–1697.

Foi,A. (2009) Clipped noisy images: heteroskedastic modeling and practical denoising. *Signal Process.*, **89**, 2609–2629.

Freeman,M. and Tukey,J. (1950) Transformations related to the angular and the square root. *Ann. Math. Stat.*, **21**, 607–611.

Guner,R. *et al*. (2011) Outcomes in patients infected with carbapenem-resistant Acinetobacter baumannii and treated with tigecycline alone or in combination therapy. *Infection* [Epub ahead of print, doi: 10.1007/s15010-011-0161-1, July 26, 2011].

Hagger,W.W. and Zhang,H. (2006) A survey of the nonlinear conjugate gradient methods. *Pac. J. Optim.*, **2**, 35–58.

Huson,D.H. *et al*. (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.

Huson,D. *et al*. (2009) Methods for comparative metagenomics. *BMC Bioinformatics*, **10**, S1—S12.

Kasuya,E. (2004) Angular transformation - another effect of different sample sizes. *Ecol. Res.*, **19**, 165–167.

Kim,D.K. and Taylor,J.M.G. (1994) Transform-both-sides approach for overdispersed binomial data when N is unobserved. *J. Am. Stat. Assoc.*, **89**, 833–845.

Knights,D. *et al*. (2010) Supervised classification of human microbiota, *FEMS Microbiol Rev.* [Epub ahead of print, doi: 10.1111/j.1574-6976, October 7, 2010].

Laubscher,N.F. (1961) On stabilizing the binomial and negative binomial variances. *J. Am. Stat. Assoc.*, **56**, 143–150.

Liu,Z. *et al*. (2010) Sparse support vector machines with Lp penalty for biomarker identification. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **7**, 100–107.

Lozupone,C. and Knights,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, 71, 8228–8235.

Mitra,S. *et al*. (2009) Visual and statistical comparison of metagenomes. *Bioinformatics*, **25**, 1849–1855.

Nagarsenker,P.B. (1984) On Bartlett's test for homogeneity of variances. *Biometrika*, **71**, 405–407.

Olivier,J. (2010) Positively slewed data: revisiting the Box-Cox transformation. *Int. J. Psychol. Res.*, **3**, 69–78.

Qin,J. *et al*. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. **464**, 59–65.

Schloss,P.D. *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.

Turnbaugh,P.J. *et al*. (2007) The human microbiome project. *Nature*, **449**, 804–810.

Vens,C. and Struyf,J. (2008) Decision trees for hierarchical multi-label classification. *Mach. Learn.*, **73**, 185–214.

Wang,Q. *et al*. (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.

White,J.R. *et al*. (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.*, **5**, e1000352

Wooley,J.C. *et al*. (2010) A primer on metagenomics. *PLoS Comput. Biol.*, **6**, e1000667.

Xu,Z. *et al*. (2010) Semi-supervised feature selection based on manifold regularization. *IEEE Trans. Neural Netw.*, **21**, 1033–1047.

Zhang,M.L. and Zhou,Z.H. (2007) Ml-knn: a lazy learning approach to multi-label learning. *Pattern Recognit.*, **40**, 2038–2048.

Zhang,H. *et al*. (2009) Human gut microbiota in obesity and after gastric bypass. *Proc. Natl Acad. Sci. USA*, **106**, 2365–2370.