Bioinformatics Advance Access published December 5, 2006

BIOINFORMATICS

Genomic Sweeping for Hypermethylated Genes

Liang Goh, Susan K. Murphy, Sayan Muhkerjee, Terrence S. Furey*

Institute for Genome Sciences & Policy, Duke University

Associate Editor: Christos Ouzounis

ABSTRACT

Motivation: Genes silenced by the aberrent methylation of nearby CpG islands can contribute to the onset or progression of cancer and represent potential biomarkers for diagnosis and prognosis. Relatively few have thus far been validated as hypermethylated in cancer among over 14,000 candidates with promoter region CpG islands. A descriptive set of genes known to be unmethylated in cancer does not exist. This lack of a negative set and a large number of candidates necessitated the development of a new approach to identify novel genes hypermethylated in cancer.

Results: We developed a general method, $cluster\_boost$, that in an imbalanced data setting predicts new minority class members given limited known samples and a large set of unlabeled samples. Synthetic datasets modeled after the hypermethylated genes data show that $cluster\_boost$ can successfully identify minority samples within unlabeled data. Using genome sequence features, $cluster\_boost$ predicted candidate hypermethylated genes among 14,000 genes of unknown status. In primary ovarian cancers, we determined the methylation status for 15 genes with different levels of support for being hypermethlyated. Results indicate $cluster\_boost$ can accurately identify novel genes hypermethylated in cancer.

Availability: Software and datasets are freely available at http://labs.genome.duke.edu/FureyLab/cluster_boost

Contact: tsfurey@duke.edu

1 INTRODUCTION

DNA methylation is an important epigenetic modification that is associated with transcriptional regulation, chromatin structure, and embryonic development (for a review, see (Robertson, 2005) and references within). The aberrent hypermethylation of CpG islands in promoter regions of key genes resulting in their transcriptional silencing has been associated with the onset and progression of human cancers (Robertson, 2005). The identification of genes consistently hypermethylated in cancer will contribute to our understanding of these diseases. These specific genes also represent potential biomarkers for the diagnoisis and prognosis of certain cancers and targets for new therapies (Laird, 2003).

Several studies have investigated computationally predicting the methylation status of CpG dinucleotides or CpG islands (Feltus *et al.*, 2003; Bhasin *et al.*, 2005; Feltus *et al.*, 2006; Bock *et al.*, 2006), but these regions were not necessarily in promoters of genes. Predictions were based on the presence of DNA sequence motifs (Bhasin *et al.*, 2005; Feltus *et al.*, 2003, 2006) and other DNA attributes such as sequence repeats and DNA structure characteristics (Bock *et al.*, 2006). Training data included either the methylation status of CpG islands in normal tissue (Bhasin *et al.*, 2005; Bock *et al.*, 2006) or in fibroblasts clones overexpressing DNMT1 (Feltus *et al.*, 2003, 2006). The use of genomic features in creating accurate classifiers has similarily been demonstrated in other epigenetic gene silencing mechanisms such as imprinted genes (Greally, 2002; Luedi *et al.*, 2005) and X-inactivated genes (Wang *et al.*, 2006).

Large scale experimentation using microarray technology (Adorjan *et al.*, 2002; Weber *et al.*, 2005; Hatada *et al.*, 2006), bead arrays (Bibikova *et al.*, 2006), and cloning and sequencing of methylated genomic sequence (Rollins *et al.*, 2006) can be used to assay the methylation status of thousands of genomic regions at a time. Despite these technological advances, we still do not have a global representation of the genomic targets of gene specific hypermethylation in cancer. We also lack understanding of why certain genes are targets for hypermethylation in disease while other genes are not.

We have developed a novel computational approach, named $cluster\_boost$, to sweep the genome for potential hypermethylated genes in cancer. Experimental evidence suggests that few genes are prone to hypermethylation (Weber *et al.*, 2005), but thousands are possible candidates. Therefore, our algorithm is specifically designed for predicting members of a minority set (hypermethylated genes) within a large unlabeled dataset (remaining genes with promoter CpG islands). The approach adapts strategies from machine learning techniques developed for imbalanced data. These data, characterized by a disproportionate distribution of samples in the positive and negative classes, have been primarily studied in areas of fraud detection, target discrimination, text classification, and computer security infringement (Kubat *et al.*, 1998; Pednault *et al.*, 2000; Japkowicz, 2003b; Chawla, 2003).

Previous studies have investigated imbalanced data in biological contexts (Choe *et al.*, 2000; Qian *et al.*, 2003; Yeo *et al.*, 2005; Plant *et al.*, 2006). Similar to classical imbalanced problems, these data contained samples for each class enabling a predictive model to be created. Our $cluster\_boost$ algorithm is the first designed not only for imbalanced data but also unlabeled data where only samples from one class are available.

2 SYSTEMS AND METHODS

To evaluate the algorithm, we created two different sets of synthetic data that are modeled after the hypermethylated genes dataset and described below. Applying $cluster\_boost$ to these datasets provides rough measures of the sensitivity and specificity of this method as well as a means to approximate parameter settings.

Using genome sequence features, we used $cluster\_boost$ to predict genes prone to hypermethylation in cancer. A subset of these were tested experimentally in primary ovarian cancers. The specific experiments performed in this validation step are detailed below.

*To whom correspondence should be addressed

2.1 Hypermethylated Genes and Sequence Features

We compiled a set of 63 genes previously reported to be hypermethylated in cancer (Supplemental Table S1). The majority of these are listed at the MD Anderson Cancer Center website (http://www.mdanderson.org/departments/methylation/) with a few additional genes extracted from literature. Each of these 63 hypermethylated genes has a promoter CpG island within 1.5Kb of its transcription start site (TSS). The CpG island annotation was taken from the UCSC Genome Browser (Kent *et al.*, 2002) and is defined as described previously (Gardiner-Garden and Frommer, 1987). Genes defined in the Known Genes annotation in the UCSC Genome Browser that have a similarly placed promoter CpG island were considered potentially hypermethylated genes. The methylation status of these 14,249 genes in cancer is not known.

For each hypermethylated and unlabeled gene, we defined a window consisting of bases 100Kb upstream and 10Kb downstream of the transcription start site. For each window, 64 DNA sequence features were extracted from the UCSC Genome Browser (Supplemental Table S2). In general, they reflect the concentration of different sequence elements, primarily repeat sequences and transcription related elements, within the window.

2.2 Synthetic Datasets

The two types of synthethic datasets created roughly modeled the hypermethylated genes dataset. For both, each sample is represented by a 64-value feature vector. There are three types of samples in these datasets: minority class samples with known labels; minority class samples with unknown labels; and majority class samples with unknown labels. The second and third categories comprise the set of unlabeled data from which we attempt to identify the minority samples. The number of samples from the minority class that are contained within the unlabeled data varied between 1% and 20% of the samples in this set.

2.2.1 Synthetic Datasets SD_1 We constructed samples for the minority and majority classes as follows. Let μ_{hi} and $\sigma_{hi}, i = 1..64$ be the mean and standard deviation for each of the 64 sequence features describing the hypermethylated genes. Let sep be a separation parameter such that for each feature i, $N(\mu_{hi}, \sigma_{hi})$ is the sampling distribution for the minority class and $N(sep * \sigma_{hi} + \mu_{hi}, \sigma_{hi})$ is the distribution for the majority class. The parameter sep, therefore, controls the degree of separability of samples from the two classes. Figure 1 displays the effect of this separation parameter. (Samples are defined by only 3 features for visualization puposes.)

Twelve datasets consisting of 14,050 samples were created. Of these, 50 were the labeled minority class samples, and the remaining 14,000 samples were considered unlabeled samples. For each dataset, a separation parameter of either 1, 2, or 3 was employed. The unlabeled data was comprised of either 1%, 5%, 10%, or 20% minority class samples and the remaining were majority class samples.

2.2.2 Synthetic Datasets SD_2 The features for SD_1 datasets were created using normal distributions, but few sequence feature in the known set of hypermethylated genes have values with this property of normality. To better simulate the hypermethylated genes data, we used feature vectors for these known hypermethylated genes to create new minority samples as follows:

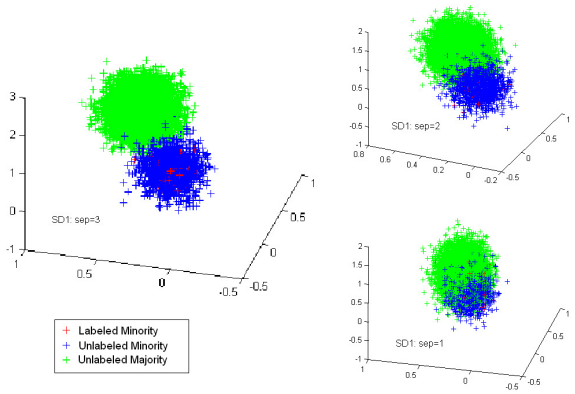

Fig. 1. Synthetic data with different separation values, $sep = 1, 2, 3$. Each sample is represented by three features. For each feature $i, i = 1..3$, $N(\mu_{hi}, \sigma_{hi})$ was the sampling distribution for the minority class and $N(sep * \sigma_{hi} + \mu_{hi}, \sigma_{hi})$ was the distribution for the majority class.

1. Randomly select 30% of the known, hypermethylated genes (minority class)
2. Within this subset, determine the five nearest neighbors of the first gene in the subset based on Euclidean distance in feature space
3. Replace 30% of the 64 feature values for the selected gene with the mean value of that feature calculated using the five nearest neighbors.
4. Repeat steps 1-3 to generate the desired number of minority samples
5. Randomly replace unlabeled samples with these new minority class samples.

A combination of the newly created minority samples and existing unlabeled data comprised the 14,249 samples in the unlabeled dataset. The number of new minority samples varied such that 1%, 5%, 10%, or 20% of this unlabeled dataset were these synthetic samples. The 63 methylated genes were used as the known minority class.

2.3 Experimental Validation

To validate predictions of hypermethylated genes, we tested for methylation in multiple primary ovarian cancers ($N = 19$ to $N = 69$, depending on the gene) using genomic DNA that was modified by sodium bisulfite as described previously (Huang *et al.*, 2006). Sodium bisulfite converts unmethylated cytosines to uracils leaving methylated cytosines unaffected. Methylation-specific (MS) PCR reactions were performed for each sample using one common primer that anneals to both methylated and unmethylated bisulfite modified DNA along with two primers that are specific to either methylated or unmethylated converted sequence. Bisulfite treated CpGenome Universal Methylated DNA (Chemicon International) was used as a positive control for methylation.

3 ALGORITHM

The discrepancy in the number of hypermethylated (minority class) vs. non-methylated (majority class) genes creates a classification

problem involving an imbalanced dataset. Several methods have been designed to compensate for imbalanced data (Cardie and Howe, 1997; Kubat *et al.*, 1998; Pednault *et al.*, 2000; Japkowicz, 2003b; Chawla, 2003; Chen *et al.*, 2004) as detailed below. The proposed techniques are aimed at ensuring that the uneven distribution of training data does not result in a biased classifier. As with traditional classifiers, these require known samples from each class to be used as training data.

Our novel $cluster\_boost$ algorithm has been designed for the general problem of predicting minority class members in this setting of imbalanced and unlabeled data. The algorithm uses a combintation of k-means clustering followed by classification using the boosting algorithm. We briefly provide some background on each of these general algorithms as well as how they are employed in our $cluster\_boost$ method.

3.1 Algorithms for Imbalanced Data

Techniques for handling imbalanced data can be categorized into supervised and unsupervised. Supervised techniques employ traditional classification algorithms, but attempt to compensate for the smaller number of samples in the minority class by either undersampling the majority class or over-sampling the minority class or both (Chawla, 2003; Japkowicz, 2003b). Cost or weight functions have also been employed to deal with this disparity in sample size (Cardie and Howe, 1997; Chen *et al.*, 2004). Optimal sampling ratios or cost functions have not been defined for the general case and have been dependent on particular datasets. The overall benefit of these methods has been a subject of debate (Chawla, 2003; Japkowicz, 2003a). Unsupervised techniques generally employ recognition-based predictors trained on samples from only one of the two classes, usually the majority class (Japkowicz *et al.*, 1995; Kubat *et al.*, 1998). In some cases, the second class is used to refine the learned class boundary.

Previous research investigated ratios of minority to majority samples in the range of 1:5 to 1:25 (Guo and Viktor, 2004; Zhang *et al.*, 2004). It has been suggested that it is not so much the imbalance but rather the inability to learn important hidden traits represented by a small number of samples in the minority or majority classes that is the cause of poor performance by standard classifiers (Japkowicz, 2003a). Thus, it is important to ensure that hidden traits in data are present during model training.

3.2 The $cluster\_boost$ Algorithm

We designed an algorithm, $cluster\_boost$, that predicts new members of a minority class from a large set of unlabeled samples. The general strategy consists of iteratively constructing imbalanced data classification problems. The unlabeled data is used to create a series of imperfect majority training sets to be used with a known minority training set. Each unlabeled sample is in a majority class training set for one experiment and in the test set for the remaining experiments. Unlabeled samples consistently classified into the minority class are predicted to be members of that class.

The algorithm is summarized in Figure 2 and consists of three main steps. First, the unlabeled data is clustered based on their feature values. These clusters should reflect aspects of the distribution of this unlabeled data in feature space. Samples are selected from these clusters in a balanced manner to create imperfect majority training sets, hopefully preserving important hidden traits in all sets. Second, a series of m classification experiments are performed using each of the majority training sets. Together, these will classify each unlabeled sample $m - 1$ times. Third, the final prediction set is determined based on the combined results of the m classification experiments.

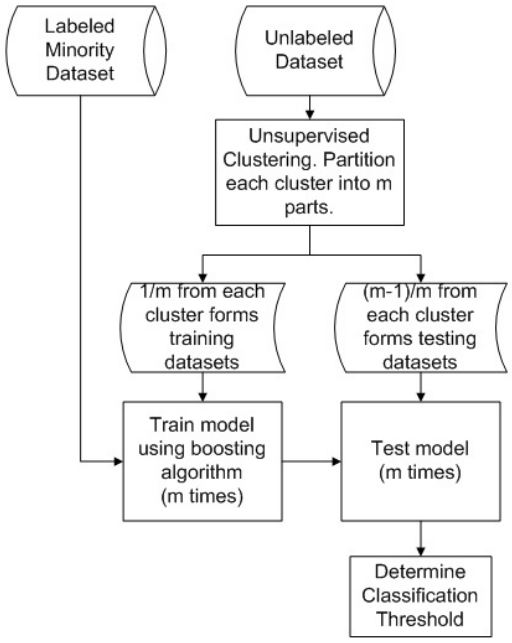

Fig. 2. Algorithm for $cluster\_boost$. Inputs are a small set of known minority samples along with a large set of unlabeled samples. The output is a set of unlabeled samples predicted to be in the minority class.

3.2.1 Clustering and Creation of Majority Class Sets The unlabeled data consists of the full set of majority class samples and some small, unknown number of samples from the minority class. From this data, we extract samples to form a series of imperfect majority training classes. In this way, we create majority training classes that are better balanced with respect to the known minority training class. At the same time, clustering allows possible hidden traits of the majority class to be preserved in each of the majority training sets.

We clustered unlabeled samples using the unsupervised k-means clustering algorithm. k-means clustering is non-deterministic and seeks to cluster samples into k clusters from k random starting points. Other clustering methods such as PAM, clara, and Clest (Dudoit and Fridlyand, 2002) were evaluated and found to be either too computationally intensive or better suited for smaller sample sizes. Initially, samples are clustered using a range of values for k to obtain a distribution of the cluster sizes. To ensure that the final clusters are robust, we repeat clustering for each k several times. We select a final k based on balancing the following properties:

1. Cluster sizes are consistently observed in every iteration.
2. Every cluster must be large enough to ensure adequate representation in all majority training sets, but not so large as to be dominant.

3. Clusters are compact as measured by the sum of distances of each sample within a cluster to its barry-centre.

Algorithm 3.1: CLUSTER_BOOST_1$(X, k_{min}, k_{max}, T_k)$

Given samples $S = \{x_1..x_P\}, x_i \in R^n$
Define
k_{min}, k_{max}: minimum and maximum number of clusters to find
T_k: number of iterations to create clusters of size k
S_{dist}: sum of distances between samples in cluster wrt barrycentre
C_{size} : size of clusters
$P_{minority}$: number of labeled minority samples
P_{freq}: commonly observed $max(C_{size})$
input: S, k_{min}, k_{max}, T_k
return:$L = max(Score_{k,i})$

for $k \leftarrow k_{min}$ **to** k_{max}
do
- **for** $i = 1$ **to** T_k
  - **do** $[S_{dist}, C_{size}, P_{freq}] = kmeans(S, k)$;
- $H_1 \leftarrow stable(C_{size})$;
- $H_2 \leftarrow C_{size} \geq P_{minority} \&\& C_{size} \leq P_{freq}$;
- $H_3 \leftarrow min(S_{dist})$;
- $Score_{k,i} = \sum_{i=1}^{3} H_i$;

The final k is chosen based on Algorithm 3.1 that scores each k based on the above properties.

Imperfect majority training sets for classification experiments are then created by first randomly dividing each cluster into m partitions, m being the number of classification experiments to be performed. Each of the m majority training sets is simply a combination of exactly one partition from each of the clusters such that each partition belongs to exactly one training set.

Algorithm 3.2: CLUSTER_BOOST_2(S, MLP, T)

Given samples $S = \{(x_1, y_1), .., (x_P, y_P)\}, x_i \in R^n$,
$y_i \in \{-1, 1\}$
Define
T: number of iterations
MLP: multi-layer perceptron w/ decision function $h(x)$
input: S, MLP, T
return:$h(x) = sign[\sum_{i=1}^{t} \alpha_i h_i(x)]$

for $i \leftarrow 0$ **to** N
do $\omega_i^0 = 1/N$;
for $t \leftarrow 0$ **to** T
do
- $h_t \leftarrow$ Call MLP with weights ω^t;
- $\epsilon_t = \sum_{j=1}^{N} \omega_j^t [y_j - h_t(x_j)]$
- $\alpha_t = log((1 - \epsilon_t)/\epsilon_t)$;
- **for** $j = 1$ **to** N
  - **do**
    - $\omega_j^{t+1} = \omega_j^t exp(-\alpha_t y_j h_t(x_j))$;
    - $Z_t = \sum_{j=1}^{N} \omega_j^{t+1}$;
    - **for** $j = 1 to N$
      - **do** $\omega_j^{t+1} = \omega_j^{t+1}/Z_t$;

3.2.2 Classification with the Boosting Algorithm Boosting is an ensemble method based on repeated presentation of difficult samples for training so that the classifiers will learn these hard samples well (Freund and Schapire, 1996). Classifiers trained using the boosting algorithm and imbalanced datasets with ratios less than 1:25 have shown robust performance (Guo and Viktor, 2004; Joshi *et al.*, 2001). This approach is well suited for imbalanced data as most discrimnation-based classifiers have a tendency to learn the majority class well but the minority class poorly. We used a modified Adaboost algorithm that adjusts weights with respect to errors between expected and actual outputs (instead of a hard-limit function for output) for a feed-forward back-propagation multi-layer perceptron (MLP), as shown in Algorithm 3.2.

3.2.3 Determination of Predictions Each sample from the unlabeled set is classified $m-1$ times in m classification experiments. In general, we expect that the more times a sample is classified into the minority class, the greater the probability it belongs to that class. Using synthetic datasets, we evaluated the accuracy of samples classified as minority samples at different classification thresholds, where the threshold denotes the minimum number of experiments a sample must be classified in the minority class. We calculated a detection acccuracy (D_{acc}) defined as the percentage of true minority samples for each threshold, 0 to $m - 1$. From this, we can estimate a minimum sensitivity and specificity at each threshold for the hypermethylated genes data.

4 IMPLEMENTATION

We first applied $cluster\_boost$ to two sets of synthetic data, SD_1 and SD_2 (see Methods), to assess the accuracy of the algorithm. We then predicted a set of novel hypermethylated genes in cancer. A small set of these predicted genes have been experimentally tested in primary ovarian cancers.

4.1 Synthetic Data SD_1

Experiments involving the SD_1 synthetic datasets demonstrated the general ability of $cluster\_boost$ to identify unlabeled minority class samples. Each dataset consisted of 50 known samples from the minority class and a set of 14,000 unlabeled samples of which a certain percentage were generated from the minority class (see Methods). SD_1 datasets evaluated the effectiveness of the algorithm on data sets with different degrees of separability as controlled by a separation parameter, sep. Also, the effect of the number of unlabeled minority samples was explored.

Each of the unlabeled datasets were initially clustered with thirty-two clusters being created each time. The smallest cluster across all experiments was 102 samples while the largest was 815 with 87% of cluster sizes being between 200 and 700 samples.

We performed $m = 20$ classifications resulting in 19 classifications of each unlabeled sample. The training set was accurately learned even with imperfect majority class training sets. Training accuracy ((TP+TN)/(TP+TN+FP+FN)[1]) was 94.5-99.1% with a sensitivity (TP/(TP+FN)) of 100% and specificities (TN/(TN+FP)) between 94.5% and 99.1%. Due to noise in the majority training class set, we did not expect 100% specificity for the training data. In fact, we found that increasing the number of unlabeled minority samples slightly decreased training set specificity reflecting the average amount of noise in this data.

[1] TP=true positives;TN=true negatives;FP=false positives;FN=false negatives

Table 1 shows average test set statistics for the 12 SD_1 datasets. It is interesting to note that the separability of the data has little effect on the results. In contrast, a higher percentage of minority samples in the unlabeled data causes a noticable degradation of sensitivity rates. This is likely due to the increase in contaminating samples in the majority training data.

Table 1. Results for $cluster\_boost$ on Synthetic Data SD_1. The first column shows the value of the separation parameter, sep. The second column shows the percentage of the unlabeled data that are minority samples. Accuracy (acc), sensitivity (sens) and specificity (spec) measures are given for test sets. For each test set, the classification threshold is shown with the corresponding number of unlabeled minority samples found out of 140 (1%), 700 (5%), 1400 (10%), and 2800 (20%) possible samples. D_{acc} for all test sets at the specified classification threshold is 100%.

sep	Unlabeled Minority (%)	Test Acc	Test Sens	Test Spec	Class Thresh	Minority Found
1	1	99.8	86.6	99.9	6	139 (99.3%)
2	1	99.8	90.3	100	6	140 (100%)
3	1	99.9	94.5	100	6	140 (100%)
1	5	97.7	56.6	100	4	660 (94.3%)
2	5	97.9	58.5	100	4	681 (97.3%)
3	5	98	60.5	100	4	698 (99.7%)
1	10	94.2	42.2	99.9	5	1125 (80.4%)
2	10	94.1	41	100	4	1155 (82.5%)
3	10	94.4	44.4	100	5	1236 (88.3%)
1	20	84.9	25.1	99.9	4	1541 (55.0%)
2	20	85.9	29.5	100	4	1908 (68.1%)
3	20	85.5	27.3	100	4	1692 (60.4%)

The high specificity rates for the test set indicates that new minority are accurately being identified. We expect the sensitivity to be less than perfect due to the relatively small number of minority training samples and the noisy majority training set. Again we see that sensitivity depends more on the percentage of unlabeled data that is from the minority set than the separability of the two classes.

We can control the number and accuracy of predictions by setting a minimum threshold for the number of times a genes is predicted to be in the minority class in all classification experiments. As expected, the more a gene is selected to be in the minority class, the greater the probability it is in that class. In table 1, we show the minimum threshold that achieves a D_{acc} of 100%. While the percentage of minority samples declines as the number of unlabeled minority samples increases, a high number of accurate predictions are still made. In an iterative way, new validated minority samples could be used to increase the labeled set and decrease the percentage of minority samples in the unlabeled set eventually leading to the identification of most or all of the minority samples.

4.2 Synthetic Data SD_2

The second set of synthetic data, SD_2, was created to more closely mimic the hypermethylated genes data. Data on the known 63 hypermethylated genes was used to create additional minority samples that were placed in the unlabeled dataset (see Methods). We varied the number of new samples that were created and repeated the entire experiment 10 times for each new sample size. Sizes of the 26 clusters created in each run ranged between 102 to 1391 samples with 85% between 200 and 1000. Again, each unlabeled gene was predicted 19 times ($m = 20$).

Table 2 displays the average results for each of the 10 iterations at each new sample size. These results only consider the synthetic data to be in the minority class, though certainly some of the real unlabeled genes also belong to this class. Therefore, the specificity is likely to be an underestimate, and the D_{acc} is probably higher as some of the "majority" class samples are, in fact, true unlabeled minority samples (hypermethylated genes).

Table 2. Results for $cluster\_boost$ on Synthetic Data SD_2. The first column shows the percentage of the unlabeled data that are synthetic minority samples. Accuracy (acc), sensitivity (sens) and specificity (spec) measures are given for test sets based on just the synthetic samples. For each test set, the best classification threshold is shown with corresponding number of unlabeled synthetic minority samples above this . Detection accuracies D_{acc} for 140 (1%), 700 (5%), 1400 (10%), and 2800 (20%) unlabeled synthetic minority samples is shown, but assumes only synthetic samples are in the minority class and is probably an underestimate. The last column shows the percent of synthetic samples classified as in the minority in at least one experiment.

Unlabel Min (%)	Test Acc	Test Sens	Test Spec	Class Thresh	Synth Found	D_{acc} (%)	Total Found (%)
1	91.2	45.0	91.6	15	24	46.2	96.2
5	90.3	38.2	93.0	15	87	75.0	93.5
10	87.9	33.3	93.9	13	196	84.8	91.0
20	81.0	25.4	94.6	13	227	89.4	84.1

Figure 3 shows the average distribution of cumulative samples found at each classification threshold for datasets with 10% synthetic unlabeled data. Similar distributions are seen for datasets with 1%, 5%, and 20% synthetic unlabeled data (Supplemental Figure SF1). The fraction of the synthetic samples at each threshold is indicated. Table 2 shows that at least 84% of synthetic samples are predicted to be in the minority class at least once.

4.3 Hypermethylated Genes Data

The results from the synthetic datasets show that $cluster\_boost$ can successfully identify unlabeled minority samples. Therefore, we applied $cluster\_boost$ to the set of 63 known hypermethylated genes and 14,249 unlabeled genes. The unlabeled data was grouped into 30 clusters ranging in size between 102 and 1389. To determine how well $cluster\_boost$ is able to learn the training data, we performed a series of training set validation tests also exploring support vector machine (SVM) and linear discriminant analysis (LDA) classifiers in addition to boosting.

First, we assessed training accuracy given the more conventional cross-validation approach. The unclustered unlabeled data was randomly split into $m = 20$ partitions. Classifiers were trained with all but one of the 20 partitions of unlabeled data and with the 63 known hypermethylated genes. We then determined how well the classifier learned this data by classifying the training data. This was repeated 20 times, each time holding out a different unlabeled partition. We also tested $cluster\_boost$ and modified $cluster\_boost$ classifiers

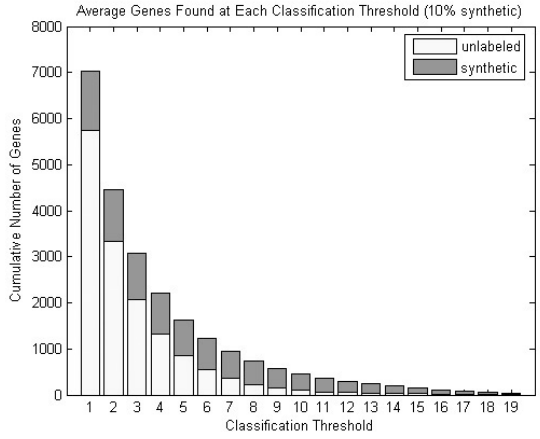

Fig. 3. Cumulative number of samples at each classification threshold for SD_2 datasets with 10% of unlabeled data synthetically created for the minority class. It is likely that some of the real unlabeled data belongs to the minority class.

substituting SVMs ($cluster\_SVM$) and LDA ($cluster\_LDA$) for boosting in the classification step. Using the 30 clusters of the unlabeled data, $m = 20$ partitions were created as described previously. Each training set consisted of a single partition of unlabeled data and the 63 known hypermethylated genes. Again, this was repeated 20 times with each unlabeled partition being in the training set once.

The results of these training set validations are shown in Table 3. In general, LDA does not perform as well as boosting and SVMs. For boosting and SVM classifiers, we see that specificity rates for more standard cross validation training data are high, but their sensitivity is low. This is likely due to the extremely imbalanced nature of the training data. Both $cluster\_boost$ and $cluster\_SVM$ have very high sensitivity and specificity with $cluster\_boost$ having a slightly higher sensitivity. Since we are most interested in accurately identifying new hypermethylated genes, the near perfect sensitivity of $cluster\_boost$ is extremely important.

Table 3. Training set validation experiments comparing boosting, SVM, and LDA classifiers using the hypermethylated gene data. In all experiments, the 63 known hypermethylated genes were part of the training along with either 19 (first three rows) or 1 (last three rows) partition of the unlabeled data.

Method	Sensitivity	Specificity	Accuracy
LDA	70.4	69.0	69.0
SVM	54.6	100	99.8
Boosting	0	100	99.8
$cluster\_LDA$	69.0	71.4	71.2
$cluster\_SVM$	94.6	99.9	99.5
$cluster\_boost$	99.7	99.7	99.7

Following the $cluster\_boost$ algorithm, we performed twenty classification experiments resulting in 19 predictions for each unlabeled gene. Figure 4 shows the cumulative number of genes found at each threshold forming a similar distribution as for the SD_2 datasets. From results using the SD_2 data (Table 2), we estimate that the 69 predictions at a classification threshold of 13 will have an accuracy of at least 50% and possibly as high as 90%. Table 4 shows the 41 genes classified as hypermethylated in at least 15 of the experiments. A complete list of genes and the number of times each was classified as being hypermethylated can be found in Supplemental Table S3.

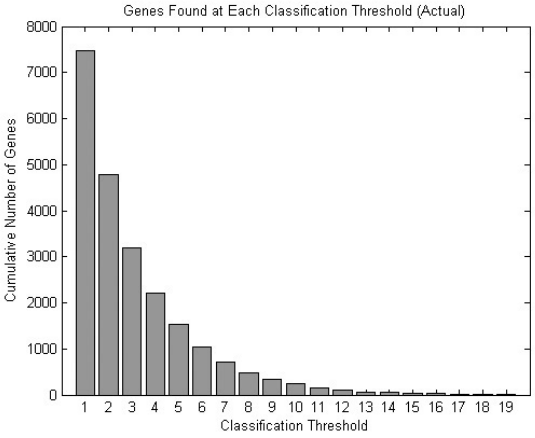

Fig. 4. Number of potentially hypermethylated genes found at each classification threshold.

Table 4. Genes classified as hypermethylated in at least 15 of 19 classification experiments. Bolded genes were found to be methylated in ovarian cancers.

SPG20	CALCB	EPS8L2	NDST2	C9orf96	RPL18A
LOC293232	HOXA6	POP7	EVX1	TUSC2	HOXA7
LOC492304	HOXA4	PL6	TUSC4	SURF2	POLR2E
FLJ30058	CKM	PDE4A	ETV6	AF116611	HYAL2
ZMYND10	TRIM32	RAVER1	DAB1	CD3EAP	NEDD8
SLC22A1LS	CHMP4A	SCT	MUCDHL	IGSF1	RPL7A
MGC19604	HOXA9	HOXA2	HOXA1	BBS5	

For 15 genes, shown in Table 5, the methylation was determined experimentally. The selection of these genes was not based solely on predictions by $cluster\_boost$. We did seek to validate genes that had been classified as hypermethylated at different thresholds to obtain a more general assessment of the accuracy of the results.

Each gene was assayed for the presence of methylation in primary ovarian cancer tissues. Genes were considered to exhibit promoter methylation when at least 5 of the analyzed cancer specimens produced an amplification product using the primer set specific to the methylated sequence (see Methods). Some genes were found methylated in less than five tissues and were considered as having an unknown methylation status. As shown in Table 5, all five genes that were classified as hypermethylated at least 5 times proved to be methylated in ovarian cancer tissues. In contrast, only two of

Table 5. Genes experimentally tested by methylation specific PCR. Reported is the classification threshold and the methylation status of each gene in primary ovarian cancers.

Gene	Description	Thresh	Methyl
SPG20	spartin	19	Yes
ZMYND10	zinc finger, MYND domain-containing 10	18	Yes
CDH4	cadherin 4, type 1 preproprotein	7	Yes
MLPH	melanophilin	6	Yes
KHDRBS2	KH domain-containing, RNA-binding signal	5	Yes
COBRA1	cofactor of BRCA1	4	?
SLC38A6	N system amino acid transporter NAT-1	4	No
CDC2	cell division cycle 2 protein isoform 1	3	No
BARD1	BRCA1 associated RING domain 1	2	No
CHEK1	CHK1 checkpoint homolog	2	No
DDX26	DEAD/H box polypeptide 26	1	Yes
IGFBP7	insulin-like growth factor binding protein 7	1	?
PLSCR1	phospholipid scramblase 1	1	?
BUB1	BUB1 budding uninhibited by benzimidazoles 1	1	No
CENPA	centromere protein A	0	Yes

the nine genes classified as hypermethylated 4 or fewer times were confirmed as methylated. The methylated genes not predicted often may contain features not well represented in the small known set. Therefore, including these in the known training set may improve accuracy and help identify new candidates not previously predicted. A more complete description of these results will be presented elsewhere.

It is interesting to note that $\approx$7000 (49%) of genes are never classified into the hypermethylated class. The results for the synthetic datasets suggest that these should contain a much lower percentage of hypermethylated genes than the full unlabeled data. Therefore, this set may be a reasonable approximation of an unmethylated dataset to be used in more traditional classification experiments, or alternatively as a subset of the unlabeled data from which to choose majority training sets for $cluster\_boost$.

5 DISCUSSION

Reasons for the aberrant acquisition of promoter methylation are still unknown. Our analyses, in addition to that by others, suggest that there is something special about the sequence context of certain promoter regions that predisposes them to becoming methylated. An initial analysis of the sequence features used to characterize genes shows that Alu repeat elements, and more generally the class SINE repeat elements, are depleted in promoter regions of hypermethylated genes. This has been reported by others for methylated CpG islands (Feltus *et al.*, 2003), though the opposite has also been claimed (Weber *et al.*, 2005). We also find a reduced amount of transcription in regions surrounding hypermethylated genes as indicated by decreased gene density and amount of sequence that is transcribed and translated. Lastly, we find an enrichment of single nucleotide polymorphisms (SNPs). A ranked list of the sequence features based on signal-to-noise ratio (SNR) calculations considering the full data set and the average rank based on SNR calculations in each of the 20 training sets is included as Supplemental Table S4.

DNA sequence features had previously been shown to be factors for CpG island methylation, though because of data availability, these studies have been confined to the analysis of small numbers of genes or CpG islands. Instead of limiting ourselves to the same constraints, we decided to develop an approach that would allow us to do a genomic sweep for hypermethylated genes. By exploring beyond traditional data mining, we hope to expand the scope to mining for unknown, thus allowing us to ask questions beyond current constraints.

While other computational methods have been developed to predict methylation status, none have directly addressed this problem in cancer. Three array-based methods have assayed for differentially methylated regions in cancer cell lines and/or tissues (Weber *et al.*, 2005; Bibikova *et al.*, 2006; Hatada *et al.*, 2006). Each provided a list of genes with evidence of their hypermethylatoin in some cancer, with 18 (Weber set) (Weber *et al.*, 2005), 36 (Bibikova set) (Bibikova *et al.*, 2006), and 400 (Hatada set)(Hatada *et al.*, 2006) unique genes in each list. These lists were not directly used in the construction of our known hypermethylated genes set, though 1, 8, and 3 genes from these lists, respectively, overlapped our known set.

We identified genes predicted in these other studies and looked for those that we also predicted in our set of 69 above the classification threshold of 13. Due to our requirement of a promoter CpG island, not all of the genes in these other prediction lists were in our unlabeled set. Only 3 of our predictions were also in the Hatada set, and none were in the other two. This lack of agreement was seen amongst the other sets as well with only one in common between the Weber and Bibikova sets (one of our known hypermethylated genes), one in common between the Weber and Hatada sets, and six (one a known hypermethylated gene) in common between the Bibikova and Hatada sets. In general, this indicates that there are likely many hypermethylated genes to be uncovered and that these discoveries will benefit from the application of several methods, both experimental and computational.

ACKNOWLEDGEMENT

SKM received support from the DoD Ovarian Cancer Research Program, award number W81XWH-05-1-0053. We gratefully acknowledge the excellent technical contributions of Yaqing Wen, Lauren R. Simel, Alison H. Gusberg and Carole Grenier.

REFERENCES

Adorjan, P., Distler, J., Lipscher, E., Model, F., Muller, J., Pelet, C., Braun, A., Florl, A. R., Gutig, D., Grabs, G., Howe, A., Kursar, M., Lesche, R., Leu, E., Lewin, A., Maier, S., Muller, V., Otto, T., Scholz, C., Schulz, W. A., Seifert, H.-H., Schwope, I., Ziebarth, H., Berlin, K., Piepenbrock, C. and Olek, A. (2002) Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res.*, **30**, e21–.

Bhasin, M., Zhang, H., Reinherz, E. L. and Reche, P. A. (2005) Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Letters*, **579**, 4302–4308.

Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E. W., Wu, B., Doucet, D., Thomas, N. J., Wang, Y., Vollmer, E., Goldmann, T., Seifart, C., Jiang, W., Barker, D. L., Chee, M. S., Floros, J. and Fan, J.-B. (2006) High-throughput DNA methylation profiling using universal bead arrays. *Genome Res.*, **16**, 383–393.

Bock, C., Paulsen, M., Tierling, S., Mikeska, T., Lengauer, T. and Walter, J. (2006) CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet.*, **2**, e26.

Cardie, C. and Howe, N. (1997) Improving minority class prediction using case-specific feature weights. In Kaufmann, M. (ed.), *Intl. Conf. on Mach. Learn.*, pp. 57–65. AAAI Press.

Chawla, N. (2003) C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In Chawla, N. (ed.), *International Conference on Machine Learning*. AAAI Press, Washington DC.

Chen, C., Liaw, A. and Breiman, L. (2004) Using random forest to learn imbalanced data. *Technical Report*.

Choe, W., Ersoy, O. K. and Bina, M. (2000) Neural network schemes for detecting rare events in human genomic DNA. *Bioinformatics*, **16**, 1062–1072.

Dudoit, S. and Fridlyand, J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.*, **3**, research0036.1–21.

Feltus, F. A., Lee, E. K., Costello, J. F., Plass, C. and Vertino, P. M. (2003) Predicting aberrant CpG island methylation. *PNAS*, **100**, 12253–12258.

Feltus, F. A., Lee, E. K., Costello, J. F., Plass, C. and Vertino, P. M. (2006) DNA motifs associated with aberrant CpG island methylation. *Genomics*, **87**, 572–579.

Freund, Y. and Schapire, R. (1996) Experiments with a new boosting algorithm. In Chawla, N. (ed.), *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 148–156. Bari, Italy.

Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.

Greally, J. M. (2002) Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *PNAS*, **99**, 327–332.

Guo, H. and Viktor, H. L. (2004) Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. *SIGKDD Explor. Newsl.*, **6**, 30–39.

Hatada, I., Fukasawa, M., Kimura, M., Morita, S., Yamada, K., Yoshikawa, T., Yamanaka, S., Endo, C., Sakurada, A., Sato, M., Kondo, T., Horii, A., Ushijima, T. and Sasaki, H. (2006) Genome-wide profiling of promoter methylation in human. *Oncogene*, **25**, 3059–3064.

Huang, Z., Wen, Y., Shandilya, R., J.R., M., Berchuck, A. and Murphy, S. (2006) High throughput detection of *m6p/igf2r* intronic hypermethylation and LOH in ovarian cancer. *Nucleic Acids Res.*, **34**, 555–63.

Japkowicz, N. (2003a) Class imbalances: are we focusing on the right issue? In Chawla, N. (ed.), *Intl. Conf. on Machine Learning*. AAAI Press, Washington DC.

Japkowicz, N. (2003b) Learning from imbalanced data sets: a comparison of various strategies. In Chawla, N. (ed.), *International Conference on Machine Learning*. AAAI Press, Washington DC.

Japkowicz, N., Myers, C. and Gluck, M. (1995) A novelty detection approch to classification. In *Proceedings of the Fourteenth Joint Conference on Artificial Intelligence*, pp. 518–523.

Joshi, M. V., VipinKumar and Agarwal, R. C. (2001) Evaluating boosting algorithms to classify rare classes: comparison and improvements. In Chawla, N. (ed.), *International Conference on Data Mining*. AAAI Press, Washington DC.

Kent, W., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., M., Z. A. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Kubat, M., Holte, R. C. and Matwin, S. (1998) Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, **30**, 195–215.

Laird, P. W. (2003) The power and the promise of DNA methylation markers. *Nature Reviews Cancer Nat Rev Cancer*, **3**, 253–266.

Luedi, P. P., Hartemink, A. J. and Jirtle, R. L. (2005) Genome-wide prediction of imprinted murine genes. *Genome Res.*, **15**, 875–884.

Pednault, E. P. D., Rosen, B. K. and Apte, C. (2000) Handling imbalanced data sets in insurance risk modeling. In Japkowicz, N. (ed.), *AAAI Workshop*, pp. 58–63. AAAI Press, Austin, Texas.

Plant, C., Bohm, C., Tilg, B. and Baumgartner, C. (2006) Enhancing instance-based classification with local density: a new algorithm for classifying unbalanced biomedical data. *Bioinformatics*, **22**, 981–988.

Qian, J., Lin, J., Luscombe, N. M., Yu, H. and Gerstein, M. (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, **19**, 1917–1926.

Robertson, K. D. (2005) DNA methylation and human disease. *Nature Reviews Genetics*, **6**, 597–610.

Rollins, R. A., Haghighi, F., Edwards, J. R., Das, R., Zhang, M. Q., Ju, J. and Bestor, T. H. (2006) Large-scale structure of genomic methylation patterns. *Genome Research*, **16**, 157–163.

Wang, Z., Willard, H. F., Muhkerjee, S. and Furey, T. S. (2006) Evidence of influence of genomic DNA sequence on human X chromosome inactivation. *PLoS Comp. Biol.* In press.

Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L. and Schubeler, D. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.*, **37**, 853–862.

Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T. and Burge, C. B. (2005) Identification and analysis of alternative splicing events conserved in human and mouse. *PNAS*, **102**, 2850–2855.

Zhang, J., Bloedorn, E., Rosen, L. and Venese, D. (2004) Learning rules from highly unbalanced data sets. In *IEEE International Conference on Data Mining*. IEEE Computer Society Press, Brighton, UK.