**Group learning for high-dimensional sparse data**

1. *Group learning for hand-written digit matrix data*

**Experiment 1: 16-digit matrix with 25% overlap between positive and negative class**

- *positive class:* 16-digit matrix composed of digits ‘7’, ‘2’, ‘8’, ‘6’ (4 images for each digit, see Fig. 1);

- *negative class:* 16-digit matrix composed of digits ‘1’, ‘2’, ‘3’, ‘4’ (4 images for each digit, see Fig. 1);

- *feature vector (for group learning)*: real-valued vector of size 784 (representing a single image (28\*28 pixel) in the digit matrix)

- number of training inputs/matrices: 20 (10 per class);

- number of validation matrices: 20 (10 per class);

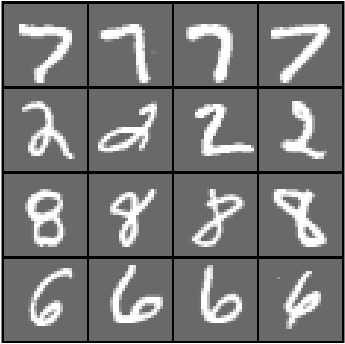
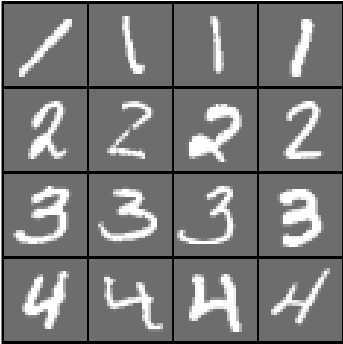
- number of test matrices: 1000 (500 per class)

**Results:**

Both SVM and Group Leaning show high SS and SP. The binary classification problem of 16-digit matrix data can be deal with by both standard SVM and Group Learning.

**Table 1.** Prediction performance for 16-digit matrix data (Experiment 1)

|  |  |  |
| --- | --- | --- |
| Method | SS | SP |
| SVM | 1.00 | 1.00 |
| Group leaning | 0.97 | 0.99 |

**Fig. 1.** Example of the 16-digit matrix in positive class (left) and negative class (right)

**Experiment 2: 16-digit matrix with 25% overlap between positive and negative class (random ordered)**

- *positive class:* 16-digit matrix composed of *random ordered* digits ‘7’, ‘2’, ‘8’, ‘6’ (4 images for each digit, see Fig. 2);

- *negative class:* 16-digit matrix composed of *random ordered* digits ‘1’, ‘2’, ‘3’, ‘4’ (4 images for each digit, see Fig. 2);

- *feature vector (for group learning)*: real-valued vector of size 784 (representing a single image (28\*28 pixel) in the digit matrix)

- number of training inputs/matrices: 20 (10 per class);

- number of validation matrices: 20 (10 per class);

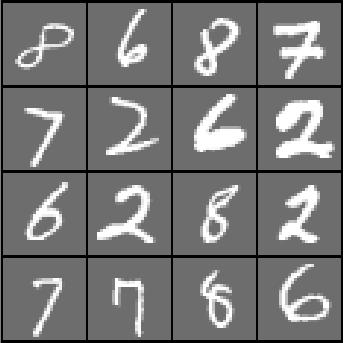
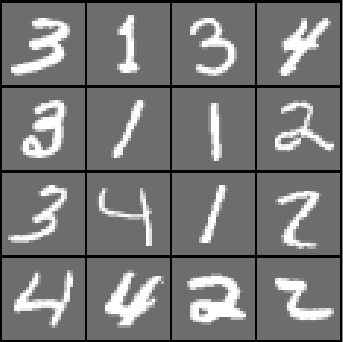
- number of test matrices: 1000 (500 per class)

**Results:**

The prediction performance of SVM for random-ordered matrix decays significantly (compared with Experiment 1). The binary classification problem becomes challenging when the digits are randomly arranged. However, Group Learning still shows good prediction performance (see Table 2). The results support the permutation invariance of Group Learning when the local spatial relationship within every single image is still valid.

**Table 2.** Prediction performance for 16-digit matrix data (Experiment 2)

|  |  |  |
| --- | --- | --- |
| Method | SS | SP |
| SVM | 0.93 | 0.77 |
| Group leaning | 1.00 | 0.90 |

**Fig. 2.** Example of the 16-digit matrix in positive class (left) and negative class (right)

**Experiment 3: 800-digit matrix with 90% overlap between positive and negative class (random ordered)**

- *positive class:* 800-digit matrix composed of a certain percentage (0.25, 0.10, 0.025, and 0.01) of digits ‘1’ and the others are even digits (‘0, 2, 4, 6, 8’) (see Fig. 3);

- *negative class:* 800-digit matrix composed of even digits (see Fig. 4);

- *feature vector (for group learning)*: real-valued vector of size 784 (representing a single image (28\*28 pixel) in the digit matrix)

- number of training inputs/matrices: 5 examples from positive class and 40 from negative class;

- number of validation matrices: 80 (40 per class);

- number of test matrices: 1000 (500 per class)

**Results:**

Group learning shows good prediction performance for the digit matrix data (see Table 3). The SS and SP remain high (SS > 0.86, SP > 0.94) for the percentages of digits ‘1’ equal or greater than 0.025. When the percentage of digits ‘1’ decreases to 0.01 (an extremely low number of digits ‘1’ in positive samples), the SS and SP start to decrease.

**Table 3.** Prediction performance for digit matrix data (Experiment 3)

|  |  |  |
| --- | --- | --- |
| Percentage of digits ‘1’ in positive samples | Sensitivity (SS) | Specificity (SP) |
| 0.25 | 1.00 | 1.00 |
| 0.10 | 1.00 | 0.98 |
| 0.025 | 0.86 | 0.94 |
| 0.01 | 0.62 | 0.78 |



**Fig. 3.** Example of the positive class matrix with 800 images which include 720 even digits (‘0, 2, 4, 6, 8’) and 80 digits ‘1’.



**Fig. 4.** Example of the negative class matrix with 800 images which are all even digits (‘0, 2, 4, 6, 8’).

**Experiment 4: 800-digit matrix (composed of half images) with 90% overlap between positive and negative class (random ordered)**

- *positive class:* digit matrix composed of 10% digits ‘1’ and the others are even digits (‘0, 2, 4, 6, 8’) (see Fig. 5);

- *negative class:* digit matrix composed of 800 even digits (see Fig. 6);

- *feature vector (for group learning)*: real-valued vector of size 784 (representing a single image (28\*28 pixel) in the digit matrix), half of the values are zero.

- number of training inputs/matrices: 5 examples from positive class and 40 from negative class;

- number of validation matrices: 80 (40 per class);

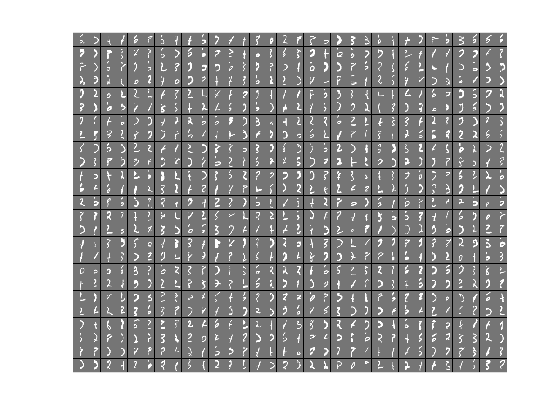
- number of test matrices: 1000 (500 per class)

**Results:**

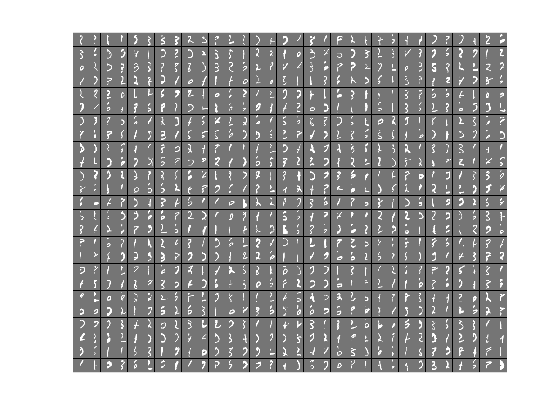
Different from the digit matrix used in Experiment 1, the binary classification problem of half-images digit matrices is difficult for a human. However, Group Learning method still shows good prediction performance (see Table 4).

**Table 4.** Prediction performance for half-images digit matrix data (Experiment 4)

|  |  |
| --- | --- |
| Sensitivity (SS) | Specificity (SP) |
| 1.00 | 0.95 |



**Fig. 5.** Example of the positive class matrix with 800 half images which include 720 even digits (‘0, 2, 4, 6, 8’) and 80 digits ‘1’.



**Fig. 6.** Example of the negative class matrix with 800 half images which are all even digits (‘0, 2, 4, 6, 8’).

1. *Statistical Invariance of Group Learning*

**Experiment 5: Group Learning using different descriptive statistics for making decision rule**

Proposed Group Learning only uses the mean value of SVM outputs of training data to perform decision rule. In Experiment 5, several descriptive statistics (including mean, std, median, 75th percentile, and 25th percentile) are used for making decision rule, as shown in Fig. 7. The experimental settings are shown below:

- *positive class:* 800-digit matrix composed of 10% digits ‘1’ and the others are even digits (‘0, 2, 4, 6, 8’) (see Fig. 3);

- *negative class:* 800-digit matrix composed of 800 even digits (see Fig. 4);

*- feature vector (for group learning)*: real-valued vector of size 784 (representing a single image (28\*28 pixel) in the digit matrix)

- number of training inputs/matrices: 5 examples from positive class and 40 from negative class;

- number of validation matrices: 80 (40 per class);

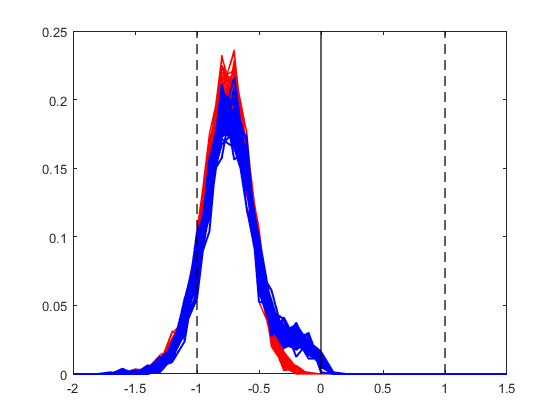
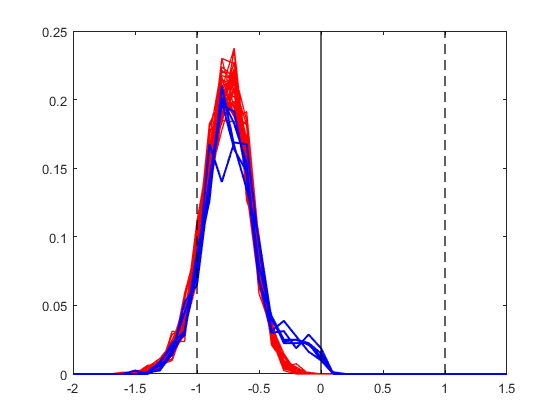
- number of test matrices: 1000 (500 per class)

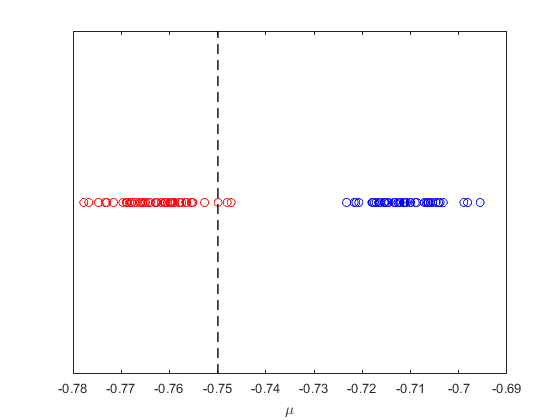
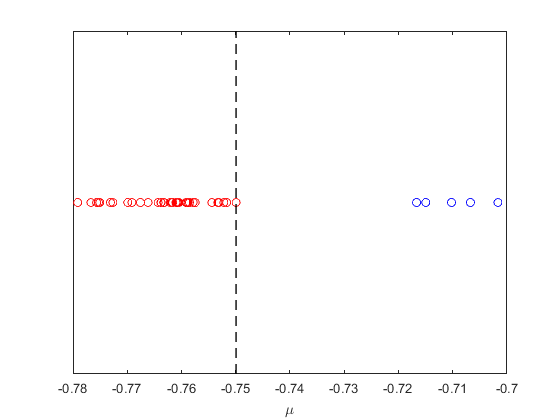
Results:

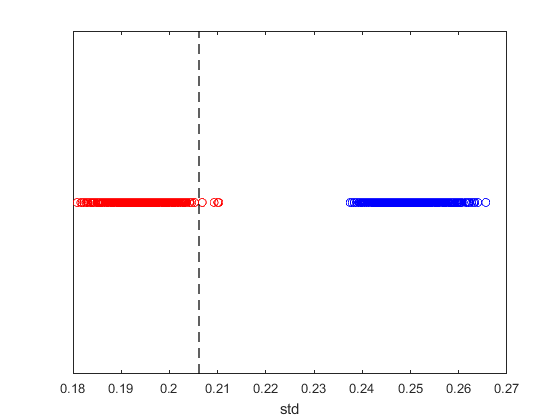
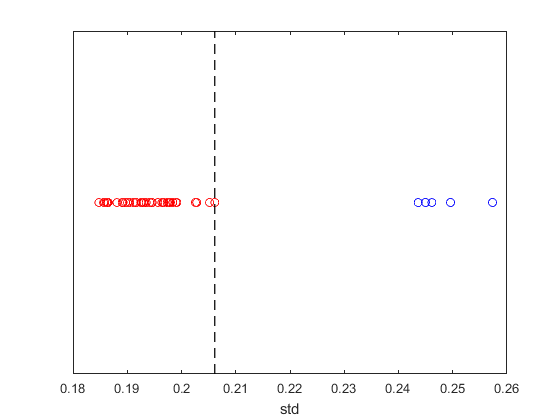
Group Learning shows strong statistical invariance when using different descriptive statistics to estimate decision threshold. The decision thresholds estimated using most of the descriptive statistics can generate similar good prediction results (see Table 5). The prediction performance decreases when using 25th percentile to estimate the decision threshold. It can be understood from the histogram of projection (in Fig. 7); that is, all left tails (of both positive and negative samples) are highly overlapping.

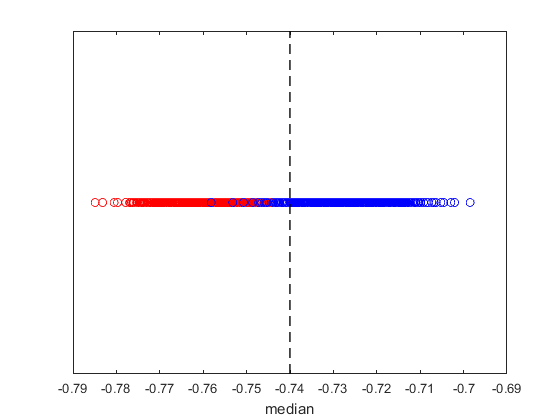
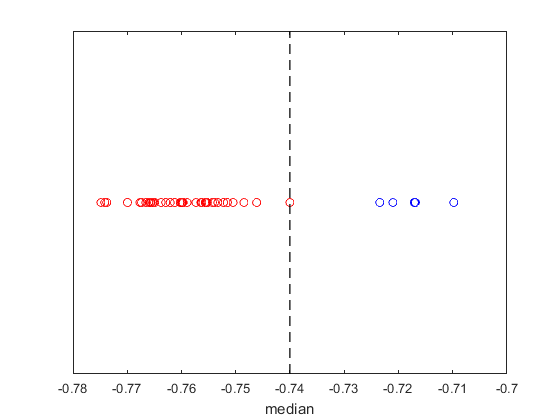
**Table 5.** Prediction performance for Group Learning using different descriptive statistics (Experiment 3)

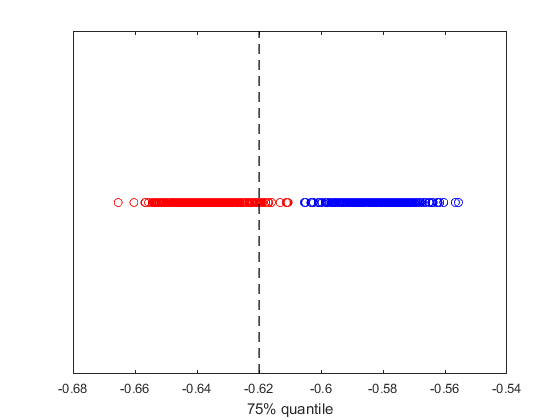
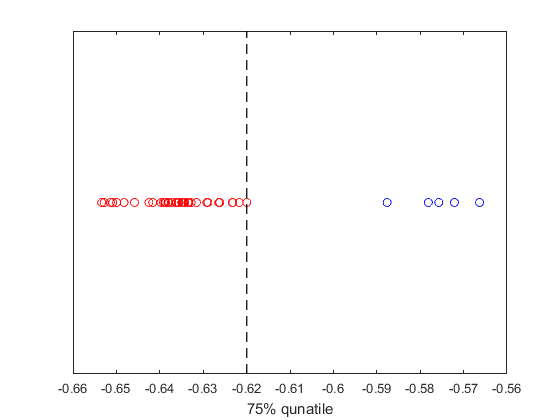
|  |  |  |
| --- | --- | --- |
| Descriptive statistic | SS | SP |
| Mean | 1.00 | 0.93 |
| STD | 1.00 | 1.00 |
| Median | 0.97 | 0.98 |
| 75th percentile | 1.00 | 0.98 |
| 25th percentile | 0.77 | 0.96 |











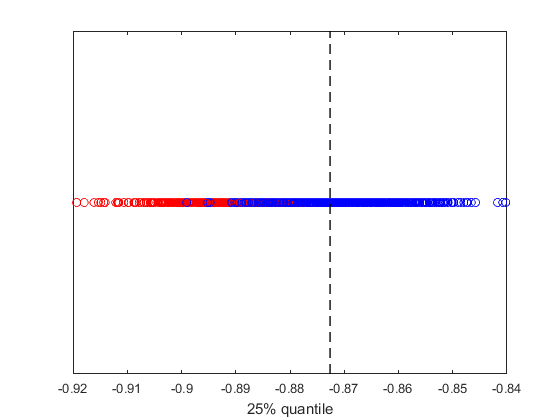
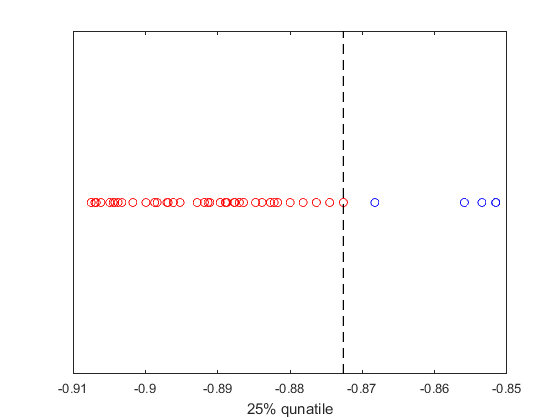


Fig. 7. Histograms of projections of training (top-left corner) and test results (top-right corner). The corresponding decision spaces and thresholds estimated from descriptive statistics are shown in the lower subfigures (from top to bottom: mean, std, median, 75th percentile, and 25th percentile).

1. *Group learning for real-life data*

**Experiment 6:** **iEEG data (detailed information in Cherkassky et al., 2019)**

*- Positive class:* 4-hr preictal iEEG segments

*- Negative class:* 4-hr interictal iEEG segments

- *iEEG data for four canine*s

**Dog L2:** 6 preictal segments and 48 interictal segments

**Dog L7:** 7 preictal segments and 56 interictal segments

**Dog M3:** 18 preictal segments and 144 interictal segments

**Dog P2:** 5 preictal segments and 40 interictal segments

*- feature vector (for group learning)*: real-valued vector of size 96 (20s iEEG window is represented by 6 features corresponding to energy in each of six standard Berger frequency bands for its 16 channels)

- The same data-analytic modeling used in (Shiao et al., 2016). Using Dog L7 dataset as an example, we have 7 experiments with 7 different models and each model is tested using its own hold-out test set (containing 1 preictal and 1 interictal). The performance indices (FN, FP) are estimated based on these seven test sets (total 7 preictal and 7 interictal).

**Results:**

Group learning shows good prediction performance for seizure prediction using iEEG data. For all four canines, all sensitivities are greater than 0.80, and specificities are greater than 0.89 (see Table 6).

**Table 6.** Prediction performance for four canines

|  |  |  |
| --- | --- | --- |
| Dog | SS | SP |
| L2 | 1.00 | 1.00 |
| L7 | 0.86 | 1.00 |
| M3 | 0.89 | 0.89 |
| P2 | 0.80 | 1.00 |

**Experiment 7:** **Gene expression data (detailed information in Ramaswamy et al., 2003, Díaz-Uriarte et al., 2006)**

*- Positive class:* metastatic adenocarcinomas

*- Negative class:* primary adenocarcinomas

- Gene expression data set: 12 samples in positive class and 64 samples in negative class, each sample is represented as 9868 features.

- *feature vector (for group learning)*: real-valued feature vector of gene expression. The size of feature vector is selected via model selection as the regularization parameter ‘*C’* of SVM. Grid search (for the optimal feature size) is applied to the range of [1250, 1400, 1650, 2000, 2450, 3300, 4900] (i.e., the original feature is split into [2, 3, 4, 5, 6, 7, 8] disjoint groups).

- The same data-analytic modeling used in (Shiao et al., 2016), which has been descripted in Experiment 6.

**Results:**

Group learning can predict most of samples in the positive class (SS > 0.75) and beyond half samples in the negative class (SP > 0.50). The high sensitivity is remarkable because of the small number of available positive samples. Since the lack of prior knowledge, the feature size is optimally selected in model selection. The feature size selected is between 1650-3300.

**Table 7.** Prediction performance for gene expression dataset (5 repeats)

|  |  |  |  |
| --- | --- | --- | --- |
| Repeat | SS | SP | Selected feature size (average) |
| 1 | 0.83 | 0.50 | ~2000 |
| 2 | 0.75 | 0.67 | ~3300 |
| 3 | 0.83 | 0.67 | ~1650 |
| 4 | 0.75 | 0.50 | ~3300 |
| 5 | 0.75 | 0.67 | ~3300 |

1. *Comparison prediction results of SVM and Group Leaning for data sets with different level of prior knowledge.*

In this section, we compare the prediction performance of Group Learning and standard SVM for the three data sets (hand-written digit matrix, iEEG, and gene expression). The Group Learning results are obtained from previous experiments. The SVM results is obtained by applying SVM classifier directly to the high dimensional sparse data. Note that the same data-analytic modeling used in (Shiao et al., 2016) is also applied for SVM experiment.

**Results:**

The comparison is shown in Table 8. It is hard for SVM to directly handle these three data sets because the data is sparse and unbalanced. However, Group Learning shows superior performance in these three data sets (compared with standard SVM method). Prior knowledge plays an important role for Group Learning. When the prior knowledge of the data set is strong (e.g., hand-written digit matrix), the Group Learning method shows brilliant performance. Through the strong prior knowledge, the feature size (for Group Learning) can be pre-determined and assured informative. For instance, the digit matrix is composed of digit images. Hence, choosing feature size equaled to a single image (28\*28=784) is reasonable. However, selecting feature size for iEEG data is not so intuitive. Fortunately, there are human experts who can help to specify the informative feature size (i.e., 20s iEEG represented as 96-dimensional feature vectors) for predicting seizures. In contrast, there is no suggested feature size of gene expression for predicting cancer type. To address this problem, we select the optimal feature size in model selection, which generates the lowest validation error. The results of Group Learning for the gene expression data is not as good as the results for the other two data sets. However, there is still a significant improvement in comparison with the conventional SVM method.

**Table 8.** Prediction performance of Group Learning applied in real data sets

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data set | Prior knowledge | SVM | | Group Learning | |
| SS | SP | SS | SP |
| Hand-written digit matrix | Strong | 0 | 1.00 | 1.00 | ~1.00 |
| iEEG | Median | Nan\* | Nan\* | 0.89 | 0.97 |
| Gene expression | Weak | 0.25 | 1.00 | 0.78 | 0.60 |

\*The feature dimension of iEEG samples (i.e., 4hr segment) is too high for SVM

Reference:

1. Cherkassky, Vladimir, Hsiang-Han Chen, and Han-Tai Shiao. "Group Learning for High-Dimensional Sparse Data." *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019.
2. Ramaswamy, Sridhar, et al. "A molecular signature of metastasis in primary solid tumors." *Nature genetics* 33.1 (2003): 49-54.
3. Díaz-Uriarte, Ramón, and Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest." *BMC bioinformatics* 7.1 (2006): 3.