A Meta-analytic approach via data integration for simultaneous Gene selection and Prediction in Alzheimer’s disease

by

Qian You, Meeta Pradhan, Amirali Jazayeri, and Samiran Ghosh

PR # 10-02

This manuscript and other can be obtained via the World Wide Web from www.math.iupui.edu

August 09, 2010
A Meta-analytic approach via data integration for simultaneous Gene selection and Prediction in Alzheimer’s disease

Qian You 1, Meeta Pradhan 2, Amirali Jazayeri 1 and Samiran Ghosh 3

1 Department of Computer and Information Science, IUPUI,
2 School of Informatics, IUPUI,
3 Department of Mathematical Sciences, IUPUI(*Corresponding Author)

Abstract

Motivation: The discovery of biomarkers, as well as classifying subjects in proper disease group is an ubiquitous task in any bioinformatics research. Over the last decade there is an exponential growth in bioinformatics research and related publications. This is also prompted by the increasing usage of publicly available data set. Unfortunately these studies are so heterogenous in terms of the analyzed biological construct, the technology used and the database organization that they can hardly be put together to do a single more powerful analysis. This type of multi-study analysis by combining several data sets is often termed as “Meta-analysis” in the literature. In this paper we present a very general framework for combining several studies under $L_1$-norm support vector machine (SVM) framework.

Results: $L_1$-norm SVM has the computational advantage over many other competing classification algorithms. It also inherits the automatic feature selection property owing to its absolute norm penalty. Proposed $L_1$-norm SVM with the multi-study generalization is first elucidated via an extensive simulation study. A new voting based strategy for choosing the optimum value of the tuning parameters, as well as feature selection is also explored. Next we apply our algorithm for classification and biomarker selection in combining several studies related to Alzheimer’s disease (AD). Using the pathway analysis software Pathway Studio, we analyzed the biomarkers with respect to their pathways and their inferences in literature with AD. As a result, our methodology identified already established as well as new pathways and biomarkers of AD. The combined data set preserve most of the top voted biomarkers from separate data sets. Also we reveal some new set of candidate biomarkers only discovered when combining several studies together using our approach.

1 Introduction

The study of meta-analysis essentially entails a simultaneous study of several individual experiments with the aim of discovering hidden pattern(s) which any single study is unable to recover. Meta-analysis is a very effective tool when underlying target disease is same. From the statistical viewpoint it often leads to higher efficacy and reduces false discovery rate by borrowing strength from seemingly different studies. In the bioinformatics context there exist many studies which are potentially targeting towards same disease but having technological as well as other dissimilarities that prevents us performing a straightforward meta-analysis. For example microarray is now quite a well understood technology, however except for very rare cases it is quite difficult to do a simultaneous analysis of two or more data sets together. Existing literature includes normalization and meta-analysis (6; 10; 11), clustering (1; 5), signature algorithms (3; 13), detection of differential expression (2; 7), scoring mechanism and Bayesian Integration (8; 12) and many others. However none of them tackle two issues simultaneously, namely classification and biomarker identification. While complete analysis of a single data set is by no means a solved problem, however, meta analysis when done in correct fashion often leads to data enrichment, further strengthening the chance of new discovery and broader understanding.

In this article our main aim is to propose a new methodology when underlying feature sets for various studies are neither mutually exclusive nor completely mutually inclusive. This is the most common scenario that one will face when we try to do a meta-analysis for a specific disease (e.g. Alzheimer’s, Prostate Cancer etc.) using a some what common technology (e.g. cDNA, microarray etc.). As mentioned earlier, in the past different mechanism have been proposed to combine different feature sets in a single predictive model. These approaches are mostly a two step process and often aimed to produce a good predictive model only. While we are not undermining the usefulness of the predictive model, it is also interesting to
see if the meta analysis produces new biomarkers. Also the two step process of model building has several variations. Many produce artificial normalized scores, while some relies on Bayesian averaging to combine several data sets in the second step to yield a single functional model. One problem with the stepwise approach is that the second step is highly dependent upon the reliability and the accuracy of the first step. In this paper we propose to build a single step model to produce data integration as well as classification via $L_1$ norm SVM. The advantage of single step model is multi-fold. The classification and data integration process is done simultaneously and as a result flow of variability as well learning process is interactive to one another. We study the performance our approach via simulation as well as on two microarray data sets on Alzheimer’s disease (AD). The biomarkers identified were further analyzed using Pathway Studio and literature to understand their significance in AD.

2 Methods

2.1 Overview of $L_1$ norm SVM

$L_1$ norm SVM is a well known binary classification technique and discussed in details by various authors (15; 16). The advantage of $L_1$ norm SVM over the popular $L_2$ norm SVM are multi folds. When a linear kernel is used for classification, owing to the nature of absolute ($L_1$) penalty it produces sparse solution. This fact makes feature selection an inbuilt task for $L_1$ norm SVM. On the other hand when a non-linear kernel is in use it indicates only a few kernel function determines the final classifier. This often results in faster evaluation and simpler model, obeying the principle of parsimony. Traditionally the simplex method has been used to solve the $L_1$ norm SVM. Recently several modification are proposed (9; 14) which make the computational evaluation a lot faster and thus ideal for solving high dimension bioinformatics problem. Owing to these flexibilities we adapt $L_1$ norm SVM as our basic tool to be used for the meta-analysis purpose. For the sake of simplicity of interpretations we stick to linear kernel through out this paper, but the developed methodology is extendable to non-linear kernels easily.

Suppose we are given a training set of input observations $x$ and output observations $y$, such that. We need to estimate a function based on $x$ which can predict $y$ well. Formally speaking, a training data set can be described as $T = \{x_i, y_i\}_{i=1}^{n}$, where the input $x_i \in \mathbb{R}^p$ and output $y_i \in C = \{-1, 1\}$ are i.i.d. pairs distributed as $(x, y)$ with probability measure $\mathbb{P}$. With a little abuse of notation, for linear kernel with non-separable classes, the corresponding optimization problem for $L_1$ SVM is given by for some $\nu > 0$,

$$\min_{\beta, \beta_0, \xi} \nu e'\xi + |\beta|, \text{ subject to } D(x\beta + e'\beta_0) + \xi \geq e', \xi \geq 0,$$

where $D = \text{Diag}(y_1, \cdots, y_n)$. The best separating hyper plane is given by, $f(z) = z'\beta + \beta_0 = 0$. The fitted classifier is $\hat{f}(z) = z'\hat{\beta} + \hat{\beta}_0$ and the classification rule is $\text{sign}(\hat{f}(z))$.

Following the descriptions of (14) we may reformulate the optimization problem in equation (1) by setting:

$$\beta_j = \beta_j^* - \beta_j^-, \beta_j^* \geq 0, \beta_j^- \geq 0 \forall j = 0, 1, \cdots, p, \quad (2)$$

which results in the linear programming problem:

$$\min_{\beta^+, \beta^-, \xi} \nu e'\xi + e'(\beta^+ - \beta^-) = \nu e'\xi + \sum_{j=1}^{p} (\beta_j^+ - \beta_j^-)$$

$$D[x(\beta^+ - \beta^-) + e'(\beta_0^+ - \beta_0^-)] \geq 0 \quad (3)$$

$$\beta^+, \beta^-, \beta_0^+, \beta_0^-, \xi \geq 0.$$ Mangasarian (14) made following proposition in solving the above problem via exact unconstrained reformulation stated below.

Proposition 2.1 The unconstrained dual exterior penalty problem for the $L_1$ norm SVM given in (3):

$$\min_{u \in \mathbb{R}^n} -\nu e'u + \frac{1}{2} \left[ \| x'Du - e \|_2^2 + \sum_{j=1}^{p} (\beta_j^+ + \beta_j^-) \right]$$

$$+ (-e'Du)^2 + \| e'\nu \|_2 + \| (\nu - e')_+ \|_2 + \| (-e')_+ \|_2, \quad (4)$$

is solvable for all $\nu > 0$. For any $e \in [0, \bar{e}]$ for some $\bar{e} > 0$ any solution $u$ of (4) generates an exact solution of the $L_1$ norm SVM classification problem given in (1) as follows:

$$\beta^+ - \beta^- = \frac{1}{e}([x'Du - e]_+ - (-x'Du - e)_+], \quad (5)$$

$$\beta_0 = \frac{1}{e} e'Du, \quad \xi = \frac{1}{e} (u - e')_+.$$ In addition to this $(\beta, \beta_0, \xi)$ minimizes:

$$\| \beta \|_2^2 + \beta_0^2 + \| \xi \|_2^2 + \| D(x\beta + e'\beta_0) + \xi - e \|_2^2$$

over the solution set of the $L_1$ SVM classification problem in (1).

2.2 SVM for Data Integration

We are now in position to describe the scheme for integrating data available from several similar studies. As described in the introduction section the motivation of this paper comes from the fact that several similar studies are available where gene expression data are produced
for a single (or similar) disease. We first propose to z-normalize each data set internally by dividing each gene by its mean and variance over all observations. This is necessary to take care of the heteroscedasticity that may present in different studies. However different studies will also possibly have different gene sets and as and a result no two studies will produce exactly the same genes even when same technology and same disease are being analyzed. For the sake of simplicity of exposition let us consider only two studies and the gene sets produced by two of them are \((x_1)_n \times p_1\) and \((x_2)_n \times p_2\). We denote by \(P_1\) and \(P_2\) as the gene identifier for the two data sets with \(|P_1| = p_1\) and \(|P_2| = p_2\). Often \(P_1 \cap P_2 \neq \emptyset\) but \(P_1 \neq P_2\) and as a result some of the genes are common to both studies, while others are specific to each studies. It is not clear how to combine both the studies to yield a better classification as well as prediction rule. Let \(|P_1 \cap P_2| = p\) and let \(q_1 = |P_2| - p\) and \(q_2 = |P_1| - p\). To make two data set appendable we need to some how match their columns. We propose to append \(q_1\) many constant columns to \(x_1\). Similar procedure can be followed for \(x_2\) so that at the end both the data sets will have \(p+q_1+q_2 (= q_1+p_1 = q_2+p_2)\) many columns. Hence common genes for two studies will be placed together and uncommon genes will produce blank sub matrix which we intend to fill by means of an arbitrary constant (say \(d\)). For the time being we consider the first data set only. Let the appended matrix is denoted as \(x_1^* = (dJ, x_1)_{n \times q_0+p_1}\), where \(J_{n \times q_0}\) is a matrix each of whose elements equals unity and \(d\) is any constant other than zero. Our claim is that such an appending will not affect the SVM classification, however will present us a plausible solution is that such an appending will not affect the SVM classification problem given in (1) as follows:

\[
\begin{align*}
\min_{\nu \in \mathbb{R}^n} & -\nu^T d + \frac{1}{2} \| \nu' (x'Du - e) \|_2^2 + \| (-x'Du - e) \|_2^2 \\
& + (\nu - \nu e)^T \| u - \nu e \|_2^2 + \| (u - \nu) \|_2^2, \\
\end{align*}
\]

is solvable for all \(\varepsilon > 0\). For any \(\varepsilon \in [0, \varepsilon]\) for some \(\varepsilon > 0\) any solution \(u \in 4\) generates an exact solution of the \(L_1\) norm SVM classification problem given in (1) as follows:

\[
\begin{align*}
\beta_0^+ - \beta^- & = \frac{1}{\varepsilon} [(x'Du - e)_+ - (-x'Du - e)_+], \\
\beta_0^+ & = \frac{1}{\varepsilon} c' Du,
\end{align*}
\]

\[
\xi = \frac{1}{\varepsilon} (u - \nu e)_+. 
\]

In addition to this \((\beta, \beta_0^+, \xi)\) minimizes:

\[
\begin{align*}
\beta & \geq \| \beta \|_2^2 + \| \beta_0^+ \|_2^2 + \| \xi \|_2^2 + \| D(x' + e' \beta_0^+) + \xi - e \|_2^2 . \\
\end{align*}
\]

Note that the right hand side of the equation (5) and (9) are exactly the same indicating the fact that SVM classifier is unaltered while appending constant columns to the original covariate matrix. The fitted classifier is also the same except for the intercept term which can be readily adjusted since the constant \(d\) is artificially chosen and well known in advance. Above invariance property of \(L_1\) norm SVM is used to combine the several data sets into a single framework in this paper.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Median PA</th>
<th>Std of PA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D_1)</td>
<td>0.8497</td>
<td>0.0318</td>
</tr>
<tr>
<td>(D_{1a})</td>
<td>0.8499</td>
<td>0.0296</td>
</tr>
<tr>
<td>(D_{2a})</td>
<td>0.7987</td>
<td>0.0818</td>
</tr>
<tr>
<td>(D_{2b})</td>
<td>0.9112</td>
<td>0.0746</td>
</tr>
<tr>
<td>(D_2)</td>
<td>0.9418</td>
<td>0.0174</td>
</tr>
</tbody>
</table>

Table 1: \(L_1\) SVM classification and 5-fold cross validation results on synthetic data sets (PA stands for prediction accuracy in each iteration)

3 Some Simulation Studies

We first use simulations to test invariance property of the proposed \(L_1\) norm SVM. The first simulation data set \(D_1\) has 63 observations for class 1 and 83 observations for class 2, and each class are generated by a normal distribution. For generating observations for class 1, we first generate 10 means \(\mu_k\) from a bivariate normal distribution \(N((-0.5, 2)', 2I)\), and then choose a mean \(\mu_k\) with probability 1/10. Observations are generated by sampling from \(N(\mu_k, I/5)\). Similarly, we generate observations for class 1 using means from \(N((1, -1)', 1)\). Therefore the first
two variables (features/dimensions) of \( D_1 \) is a mixture of normal clusters from two classes. Next for each observation we deliberately add two more dimensions and fill them with white noise. This essentially means only the first two dimensions of \( D_1 \) should be selected by L1 SVM for the binary classification. Now to show that appending features with constant values does not affect classification (as described in section 2.2), we append two more dimensions with constant numbers in the data set \( D_1 \) to create \( D_{1a} \) (hence \( n_1 = 146, p_1 = 4, \) but \( n_{1a} = 146, p_{1a} = 6 \)).

We choose constant \( d = 2 \), however any constant other than zero can be chosen. We then run L1 norm SVM on \( D_1 \) and \( D_{1a} \) individually. Both runs use the 5-fold cross validation over 1000 iterations. The same validation setting is used for assessing the performance of L1 norm SVM in the following studies. Results of the classification are shown in the first two rows in Table 1. “Median PA” stands for median of prediction accuracies in 1000 iterations, while “Std of PA” stands for standard deviation of prediction accuracy in 1000 iterations. As can be observed, the results of both the data sets \( D_1 \) and \( D_{1a} \) are almost identical. It is consistent with our theoretical result in section 2.2. Results of the feature selection performance of L1 norm SVM on \( D_1 \), shown in Figure 1 is assessed by comparing the estimated and selected features (\( \hat{\beta} \)) over 1000 iterations. In Figure 1 (a) and (c), represent the kernel density plots for those features in \( D_1 \) generated from the bivariate normal distribution. Their profiles do not contain central zero in the x-axis. This means that corresponding 95% confidence interval for the significance of each feature will also not contain zero. Thus these two features are significant for classifying \( D_1 \), which is consistent with the prior knowledge. Figure 1 (b) and (d) represent kernel density plots for the two features generated from white noise. Both of their profiles cover central zero in the x-axis, so the hypothesis of non-significance is accepted. Results of the feature selection from \( D_{1a} \) are shown in Figure 2. The kernel density plots in (a), (b), (c) and (d) represent the same features as in \( D_1 \), and they are almost identical to those in Figure 1. This shows that the performance of the four features is invariant before and after appending constant valued features. Figure 2 (e) and (f) represent the two appended constant valued features, and their profile contains zero in the x-axis, indicating their non-significance. From comparing results of \( D_1 \) and \( D_{1a} \), we can conclude that appending constant valued feature/s barely affect the performance of classification and feature selection of L1 norm SVM. Next we generate the second simulation data set to evaluate the performance of L1 SVM when we combine two data sets having some of the features in common. This simulation study is adhering to the procedure described in Section 2.2. We named the data sets as \( D_{2a} \) and \( D_{2b} \). The former set is created by sampling from \( N((-1, 1)^T, 2I) \) for class 1 and \( N((1, -0.5)^T, 2I) \) for class -1, while the latter set is created by sampling from \( N((-2, 2)^T, 2I) \) for class 1 and \( N((1, 0)^T, 2I) \) for class -1. The exact sampling procedure is same as described earlier in \( D_1 \) creation. We add one more column in each data set which is nothing but white noise. Hence \( D_{2a} \), \( n_{2a} = 88 \) and \( p_{2a} = 3 \) with 52.2% in class 1. For \( D_{2b} \), \( n_{2b} = 101 \) and \( p_{2b} = 3 \) with 42.5% in class 1. Clearly two data sets have one common informative feature (with the same mean of the first feature) and one uncommon but informative feature (different mean of the second feature). Each has one more feature which are purely non-informative white noise. Then we appended \( D_{2a} \) and \( D_{2b} \) to create \( D_2 \) having \( p_2 = 5 \) and \( n_2 = 88 + 101 \) with \( d = 2 \) following the procedure as described in Section 2.2. The classification result for \( D_{2a} \), \( D_{2b} \) and \( D_2 \) on L1 norm SVM is presented in Table 1. Clearly the performance of the appended data set is better than the individual performance of each data set. We assess the feature selection property in classifying \( D_2 \) as well. However due to the space limitation the kernel density plots are not provided here and can be obtained.

![Figure 1: Kernel density plots for the feature coefficients obtained from \( D_1 \) over 1000 iterations. (a) and (c) are for features generated from a mixture of bivariate normal distributions; (b) and (d) are for the two white noise features. Two features in (a) and (c) are significant as the density plot does not contain zero in the x-axis.](image)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Median PA</th>
<th>Std. of PA</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_A )</td>
<td>0.76</td>
<td>0.0642</td>
</tr>
<tr>
<td>( D_B )</td>
<td>0.54</td>
<td>0.0926</td>
</tr>
<tr>
<td>( D_C )</td>
<td>0.63</td>
<td>0.0812</td>
</tr>
</tbody>
</table>

Table 2: L1 SVM classification and 5-fold cross validation results on gene expression data set.
as a supplementary material on request. Three features are found to be significant as the kernel density plot do not contain zero in x-axis. These include one common feature of $D_{2a}$ and $D_{2b}$ and one distinct but informative feature of each. Above simulation studies show the invariance property of $L_1$ norm SVM’s feature selection and classification performance, against appending artificial constant valued feature. They also indicate that $L_1$ norm SVM’s improved performance by aggregating more than one data set with improved performance.

We should mention that we do not aim to provide here a solution for the implicit assumptions that any meta-analysis study makes, such as publication bias and varying quality of the studies. Rather the aim of this paper is to provide a one step classification and automatic feature selection technique by aggregating more than one data set with improved performance.

4 Studies on Microarray Expressions

4.1 Microarray data sets for Alzheimer’s Diseases

Now we apply the same methodology to study the feature selection and classification by combining two Alzheimer’s Diseases microarray expressions data sets produced by GeneChip Operating Software (previously known as Affymatrix MAS). To map prob sets expressions to the gene expressions, we first use the absence/presence call threshold 20% as used in the original paper to filter out unqualified genes, then map probeset ids and their expressions to Entrez Gene Symbols and expressions. The first gene expression data has 23 samples, from a 161 sample data set of AD and normal aged brain (GSE5281; Liang,2008). Out of 23, 13 are control samples and 10 are samples from AD hippocampus region. 10491 genes and their expressions are obtained. We use $D_A (n_{DA} = 23$ and $p_{DA} = 10491)$ to denote the first microarray data set. The second data set has 31 samples from endocirnal cortex (GSE1297; 4). Among these, 9 are controls, 7 are incipient stage samples, 8 are moderate stage samples and 7 are severe stage samples. For the binary classification problem, we use all 22 samples of differing AD stages as positive (or disease) samples. 7775 genes and their expressions are then obtained from the probsets of the second data set. We denote the second microarray data set as $D_B (n_{DB} = 31$ and $p_{DB} = 7775)$.

4.2 $L_1$ norm SVM Result

Referring to the appending scheme of $D_2$, we integrate $D_A (n_{DA} = 23$, $p_{DA} = 10491)$ and $D_B (n_{DB} = 31$, $p_{DB} = 7775)$. They share 5531 common features (genes). There are other 7206 features either present in $D_A$ or $D_B$. In the integration we fill in constant number $d$ (here we choose $d = 2$) for the missing value of those features. As a result we have the data set $D_C (n_{DC} = 54$ and $p_{DC} = 12735)$. Then we apply $L_1$ norm SVM on $D_A$, $D_B$ and $D_C$ separately. The classification results are presented in Table 2. $D_A$ has the highest median PA, and
The integration approach is two folds. First, from the invariance principle or feature selection even when the problem is high dimensional, it may/will potentially uncover many new features which are undiscovered in doing feature selection for individual data sets. Second, after integration the total number of correct classification is observed for $D_C$, followed by $D_A$ and $D_B$.

<table>
<thead>
<tr>
<th>Region</th>
<th>#</th>
<th>Pathway</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>26</td>
<td>Cell cycle Regulation</td>
<td>3.84E-157</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B cell Activation</td>
<td>2.57E-88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Actin signaling</td>
<td>5.44E-66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Notch Pathway</td>
<td>1.57E-59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hedgehog Pathway</td>
<td>1.86E-42</td>
</tr>
<tr>
<td>B</td>
<td>13</td>
<td>Cell cycle Regulation</td>
<td>6.61354E-317</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hedgehog Pathway</td>
<td>1.55E-50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Guanylat Cyclase Pathway</td>
<td>1.30E-44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Atlas of signaling</td>
<td>3.43E-14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Insulin Action</td>
<td>1.46E-12</td>
</tr>
<tr>
<td>AB</td>
<td>9</td>
<td>Hedgehog Pathway</td>
<td>1.74E-75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Apoptosis Regulation</td>
<td>1.08E-37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Axon Guidance</td>
<td>8.08E-19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cell cycle Regulation</td>
<td>9.52E-14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Insulin Action</td>
<td>5.87E-11</td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>Cell cycle Regulation</td>
<td>0.000146</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CREB/ELK-SRF signaling</td>
<td>0.017727</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B cell Activation</td>
<td>0.021353</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hedgehog Pathway</td>
<td>0.034861</td>
</tr>
<tr>
<td>AC∪BC</td>
<td>8</td>
<td>Cell Cycle Regulation</td>
<td>0.007405</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B cell Activation</td>
<td>0.011019</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Notch signaling</td>
<td>0.022544</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NF-KB signaling</td>
<td>0.022544</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RB1/E2F signaling</td>
<td>0.043612</td>
</tr>
</tbody>
</table>

Table 3: Top scoring pathways of each region

<table>
<thead>
<tr>
<th>Region</th>
<th>Hub Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>MAPK8, EP300, MAPK14, CTNNB1, RB1, RAF-1, IGF1R, PTPN11, TBP, FYN</td>
</tr>
<tr>
<td>B</td>
<td>EP300, MAPK8, CTNNB1, RB1, RAF-1, IGF1R, PTPN11, TBP, FYN, TGFA</td>
</tr>
<tr>
<td>AB</td>
<td>ASB8, U2AF1, UGDH, SMFD1, FAM48A</td>
</tr>
<tr>
<td>AC∪BC</td>
<td>MAPK3, EP300, MAPK8, MAPK14, HSPA1A, ELK1, CBL, TBP, RB1, FYN</td>
</tr>
<tr>
<td>N</td>
<td>CREB1, BRCA1, GSK3B, ATR, ATM, FAS, MAPT, CCNB1, VIP, RAD17</td>
</tr>
</tbody>
</table>

Table 4: Hub node Biomarkers identified in the above regions

4.3 Feature Selection Results

One of the advantages of $L_1$ norm SVM is that it can do feature selection even when the problem is high dimensional or $n << p$. The additional advantage of our data integration approach is two folds. First, from the invariance principle $L_1$ SVM will preserve the highly significant features in individual data sets. Second, after integration it may/will potentially uncover many new features which are undiscovered in doing feature selection for individual data sets. This is a serious advantage considering too many genes present in each study with very low signal strength. Also note that, given the large number of genes, we also use a voting mechanism to stabilize features selected by $L_1$ SVM over 1000 runs. The features selected in each run are not always identical and depends upon the tuning parameter in SVM. This is typically due to two reasons. First, the data set is often noisy, and second due the over fitting. Both of them are unavoidable in real life situation especially when the number of genes ($p$) are much larger than the number of observations ($n$). Hence we took a conservative approach and out of 1000 runs we consider only those genes which are selected at least 10% of the time. As a result we have three gene sets derived from $D_A$, $D_B$ and $D_C$, namely A, B and C. We divide the three gene sets into regions as represented in figure 3. Region AB represents the features common between A and B; Region BC represents the features common between B and C; Region AC represents the features common between A and C; Region AC∪BC is union between AC and BC; Region N represents features in C but not in A or B. In the following context, we would like to validate that genes selected by $L_1$ SVM and voting scheme. We also intend to show that our meta-analytical approach can preserve important genes identified by single data set study, and can uncover genes that cannot be achieved by single data sets study. We did an in depth analysis of features of each region via Pathway Studio(Nikitin,2003).

5 Functional Inference of the Biomarkers Identified

Pathway Studio is an interactive database manually created from literature publications. It is mainly used for functional analysis and validation of experimental or literature data. Differentially expressed gene data from microarray experiments can be directly uploaded on the Pathway Studio and analyzed to identify the relevant...
genes for any given complex diseases. In our case we are using it to study/validate the biomarkers that were identified using the methodology described above. We next describe findings in each region of the Figure 3. Region A: 1049 candidate biomarkers are identified in this region, including 83 transcription factor, 22 receptor ligand and 944 are identified as proteins. Region B: 775 candidate biomarkers are identified in this region, including 85 transcription factors, 11 receptors ligands and 679 are identified as proteins. Region AB: 63 candidate biomarkers are identified in this region, including 3 transcription factor and 60 are identified proteins. Region N: 48 candidate biomarkers are identified in this region, including 3 transcription factors and 14 receptors. Region AC∪BC: 813 candidate biomarkers are identified in this region, including 87 transcription factor and 726 proteins.

5.1 Significance of Pathways Associated with Biomarkers in each region

To validate the above biomarkers we first analyze them along with the pathways they belong to. The top scoring pathways of each region with respect to their significant p-values identified by Pathway Studio are given in the Table 3. From Table 3 we identify cell-cycle regulation pathway in five regions, followed by Hedgehog pathway in four regions, B cell activation in three regions, Notch signaling and Insulin pathway in two regions and the remaining eight pathways are unique to their regions.

Using literature these pathways are further validated with respect to their significance in AD. A brief discussion of few of the pathways are given next. Cell Cycle Regulation Pathway: cell-cycle regulation initiates and mediates the neurodegenerative process. Cell cycle regulation is one of the major pathways identified along with Wnt signaling for drug targets for treating Alzheimer (17). In this pathway we identified CDK6, CDC37 etc by our methodology. CDK6 and CDC37 are identified in AD. CDK6 is already a proven biomarker. Hedgehog Pathway: the SHH gene and presenilin genes promote cell proliferation and apoptosis, which are involved in Hedgehog pathway. Both SHH gene and presenilin genes are identified by our methodology in this pathway. In addition we identified SP-1, a zinc-finger transcription factor as biomarkers in this pathway. Both presenilin and SP-1 zinc finger transcription factors are identified as significant biological entities in AD.

From Table 3 we analyze that each region identifies some unique pathways with significant p-values. For example, Apoptosis regulation pathway is identified in Region AB, NF-kB signaling pathway in Region AC∪BC etc. Both these pathways are already established in literature as significant with AD. Actin Cytoskeleton Pathway is identified in Region A. It is identified recently as playing an important role in AD, suggesting the role of involvement of structural proteins in AD (29). CREB signaling pathway is identified as significant in Region N. CREB protein is responsible for memory of brain. Build-up of β amyloid is associated with AD and this build-up reduces the activity of the central region of human brain associated with memory (28).

From this brief analysis we can postulate that pathways associated with biomarkers in each region by our methodology are significant with respect to AD. Two important pathways, among others, the cell-cycle regulation and Hedgehog pathways are both identified in region A and B. This indicates that the candidate biomarkers from the combined data set Dc preserve the major pathway information which can be achieved by genes shared by data set A and B. In addition, CREB pathway is uniquely identified in region N which contains biomarkers selected by L1 SVM when microarrays are combined. Genes in the region AB, however, can not recover this pathway.

5.2 Significance of hub nodes biomarker with AD

The scope of this paper limits the study of biomarkers to the hub node with maximum connectivity. Table 4 gives the hub node profile of the biomarkers in different regions. We compare biomarkers identified in each region. It was observed that certain biomarkers of region AB are identified in A and B but not as hubnodes. Hence they are not given in the Table 4. In addition we identify some hub node candidate biomarkers shared by more than one regions. For example EP300 by regions: MAPK8, RB and TBP; FYN in by regions, MAPK14, CTNNB1, RAF1 and IGF1R; and PTPN11 by two regions. Due to limited scope it is not possible to analyze all the hub nodes biomarkers in table 4. For the exposition purpose we analyze a few of these to validate their significance in AD using literature. In the following paragraph we discussed the hub nodes shared by two or more regions except for region N. For hub nodes in region N, we discuss them in the next section separately. From Table 4 we identify EP300, MAPK8, RB1, TBP, Fyn as hub node in three regions, region A, B and AC∪BC. IGF1R and PTPN11 are hub nodes common in region A and B. The transcriptional co-activator EP300 is the regulator of many genes. EP300 was identified as important gene in the dense network of AD (18). MAPK8 is important candidate gene in AD (18). MAPK8 is asso-
associated with phosphorylation of AβPP gene identified in AD in the JNK cascades. RB1 regulates the neuroprotective activity of CXCL12. CXCL12 protects neurons from apoptosis induced by RB loss. CXCL12 is mediated by its receptor CXCR4, and CXCR4 stimulation is important to prevent neurodegenerative and neuroinflammatory conditions (19). TBP, the TATA-binding protein, is associated with Huntington Disease. Soluble TBP is identified in AD, though it doesn’t directly co-relate with accumulation of Tau or β-amyloid structure. But, polyQ repeats identified in TBP are associated with AD and hence, one can postulate the significance of TBP in AD (21). Fyn plays an important role in neurodevelopment and is associated with schizophrenia, alcoholism and epilepsy. In AD, it is associated with phosphorylated tau and Aβ (22). IGF1R is a growth hormone that stimulates growth and its activity falls with aging. Unusual distribution of IGF1R expression levels are identified in AD brain especially in the regions of neurofibrillary tangles and these levels have a significant effect on the IGF-1R signaling pathway (20). No direct correlation of PTPN11 with AD is reported in literature. PTPN11 mutations are associated with RAS and MEK pathways which are significant in AD, hence we can hypothesize that PTPN11 play an important role in AD.

In summary the region A∪B∪C share half of the hub nodes identified in region A and B. Those hub nodes are of high significance to AD. It shows again gene set C preserves major hub nodes information of gene set A and B.

5.3 Significance of hub nodes biomarker of Region N

From Table 4 we observed that the candidate biomarkers identified in region N are very unique i.e. these biomarkers are not identified in any other regions. CREB1, BRCA1, GSK3B, ATR, ATM, FAS, MAPT, CCNB1, VIP, RAD17 are the hub node biomarkers of this region. CREB1 gene encodes regulatory protein CREB that plays an important role in human brain. CREB signaling pathways have direct interactions with Aβ and these pathways are already identified to have potential for drug treatment (23). BRCA1 is a regulator of cell-cycle and is identified in neurofibrillary tangles of AD in Brain (24). ATM is associated with many neurodegenerative disorders and is associated with DNA damage. Though AD is associated with DNA repair defects, but direct correlation of ATM and AD is not identified in literature. Recent studies have identified association of FAS gene with AD by modulating apoptosis and neuronal loss. FAS gene SNPs are identified with AD progression (25). CCNB1 (cyclin B1) is accumulated in degenerating brain AD (26). From this brief analysis we can conclude that biomarkers found region N are significant with respect to AD. These candidate biomarkers, however, are not voted as the top 10 percent when L1 SVM is running on any of the single data sets.

Till this far we can conclude that candidate biomarkers identified in table 4 and analyzed in section 5.2 and 5.3 are significant to AD. Our methodology identified nodes like PTPNp11, TBP, GSK3B, ATR, ATM and RAD17 whose direct correlation with AD is not reported and these can further be validated in the wet labs. And we observed in section 5.2 that region A∪BC in gene set C preserves the major hub nodes found in gene set A or gene set B. Also region N offers many hub node biomarkers which can only be identified in the combined data set using our meta-analytical approach.

Aside from the hub node biomarker we have discussed, we have done a comprehensive analysis on the genes in different regions. We show only a small portion of the analyzed candidate biomarkers in Table 5. We list a few genes sharing by three regions, a few by two regions and a few by one region only. Those genes are found both as hub nodes with good connectivity and in the significant pathways of the region containing them. Genes marked with Y* are directly mentioned in certain literatures as biomarkers of neuro-degenerative diseases. Genes marked with Y are identified in our methodologies and need further wet lab experiments to validate its relevance with AD.

6 Discussion

Here we present a methodology to simultaneously classifying microarrays and selecting biomarkers on several similar data sets. Though we presented the integration result for only two data sets at a time, it can be easily extended to integrate more than two data sets. Our scheme is rather simple and based on well understood L1 SVM. The z-normalization done for pre-processing is fairly standard, and resultant biomarker selected are readily interpretable. Our simulation results show the exact feature secretion as well as boosted performance on the integrated data set, when we have a clear understanding of the data generating mechanism. For the sake of space here, we are not presenting similar results for more than two data sets, however our experience shows we get similar performance boosting there too. We have analyzed two microarray data sets on Alzheimer’s disease in depth. Strong biological correlations of most the biomarkers and their associated pathways with AD were identified using Pathway Studio and literature. In addition to already established pathways and biomarkers, we identified new pathways and biomarkers that has poten-
combined data sets also uncover unique pathways such as the CREB pathway which a single study can not discover. Also our proposed meta-analytic approach can be easily generalized to more than two data sets.

References


