A HYPERGRAPH-BASED LEARNING ALGORITHM FOR CLASSIFYING ARRAYCGH DATA WITH SPATIAL PRIOR

Ze Tian, TaeHyun Hwang and Rui Kuang

Department of Computer Science and Engineering, University of Minnesota, Twin Cities, 200 Union Street SE, Minneapolis, MN 55455, USA,
Correspondence: kuang@cs.umn.edu

ABSTRACT

Array-based comparative genomic hybridization (array-CGH) has been used to detect DNA copy number variations at genome scale for molecular diagnosis and prognosis of cancer. A special property of arrayCGH data is that, among the spot-intensity variables in the arrayCGH data, there are spatial relations introduced by the layout of the probes along the chromosomes. Standard classification algorithms are not capable of capturing the spatial relations for accurate cancer classification or biomarker identification from the arrayCGH data. We introduce a hypergraph based learning algorithm to classify arrayCGH data with spatial priors modeled as correlations among variables for cancer classification and biomarker identification. In the experiments, we show that, by incorporating the spatial relations among the spots as prior, our algorithm is more accurate than other baseline algorithms on a bladder cancer array-CGH data. Furthermore, some discriminative regions identified by our algorithm contain genomic elements that are cancer-relevant.

1. INTRODUCTION

DNA copy-number variations (CNVs) are the events of amplification or deletion of a large segment of DNA on chromosomes. CNVs play an important role in tumorigenesis. Previous studies have shown a close relationship between the CNV patterns and cancer cell lineages [1][2]. It is believed that characterization of these DNA copy-number changes is beneficial for both the basic understanding of cancer and its diagnosis [3]. With the advent of arrayCGH technology, high-resolution arrayCGH datasets were generated in many studies of cancers such as brain, prostate, colon, pancreatic and lung cancers [4][5][6][7]. The array-CGH data was used to discriminate healthy patients from cancer patients and classify patients of different cancer subtypes. Thus, arrayCGH data is considered as a new source of biomarkers that provide important information of candidate cancer loci for the classification of patients and discovery of molecular mechanisms of cancers.

Previous studies [8] and [9] have used DNA copy numbers for cancer discriminant analysis. These methods ignore the spatial correlation among sampling spots in array-CGH data, and thus suffer from two drawbacks: the accuracy is low and the classifier lacks interpretability. Recently [10] proposed a SVM variant to take the spatial information into consideration for classification of CGH/arrayCGH data. [10] proposed a supervised classification method which is a variant of $L_1$-SVM. By incorporating the biological specificities of DNA copy number variations along the genome as prior knowledge, they introduced new constraints to $L_1$-SVM, which leads to a new algorithm called fused SVM. They demonstrated that the new algorithm improved both the classification accuracy and the identification of discriminative regions of the genome on two cancer datasets.

In this paper, we introduce a hypergraph-based iterative learning algorithm called HyperPrior to integrate CNVs with more general spatial prior information for robust cancer outcome prediction and biomarker identification. HyperPrior using spatial relations as a priori information is a sibling of the method proposed by us in [11] for integrating gene expressions and protein-protein interactions. HyperPrior algorithm formulates an optimization problem as learning labels and hyperedge weights together with the assignment of edge weights constrained by the neighbor relation between spot regions on chromosomes. The neighbor relation is based on the fact that adjacent sampling spots are more likely to be involved in the same DNA insertion or deletion event. These properties of HyperPrior algorithm promise to improve prediction accuracy and provide more robust identification of discriminative spots of CNVs. Furthermore, the resulted weights on the spots can be used to discover highly weighted DNA regions on chromosomes, which might also suggest important gene amplification and deletion events related to the genetic mechanism of cancer.

2. HYPERGRAPH AND SPATIAL PRIOR

A hypergraph is a special graph which contains hyperedges. In a simple graph, each edge connects a pair of vertices,
but in a hypergraph each edge can connect arbitrary number of vertices in the graph. Let \( V = \{v_1, v_2, \ldots, v_{|V|}\} \) be a set of vertices and \( E = \{e_1, e_2, \ldots, e_{|E|}\} \) be a set of edges defined on \( V \): for any hyperedge \( e \in E, e = \{v_1, v_2, \ldots, v_{|e|}\} \), where \( \{v_1, v_2, \ldots, v_{|e|}\} \) is a subset of \( V \). A hyperedge \( e \) and a vertex \( v \) are called incident if \( v \in e \). A non-negative real number (a weight) can be assigned to each hyperedge by a function \( w \) (\( w \) can also be defined as a vector variable and we will use both notations interchangeably). The vertex set \( V \), hyperedge set \( E \) and the weight function \( w \) fully defines a weighted hypergraph denoted by \( G(V, E, w) \). The incidence matrix \( H \) for hypergraph \( G(V, E, w) \) is a \( |V| \times |E| \) matrix with elements defined as \( h(v, e) = 1 \) (or a real positive value if \( H \) is weighted) when \( v \in e \) and 0 otherwise. The degree of a vertex \( v \) is defined as \( d(v) = \sum_{e \in E} h(v, e) \), which is the (weighted) sum of the weights of the hyperedges incident with \( v \). The degree of a hyperedge \( e \) is defined as \( d(e) = \sum_{v \in E} h(v, e) \), which is the number of vertices incident with \( e \). Fig. 1 shows how a hypergraph is built from an arrayCGH dataset. In arrayCGH data, each spot is assigned a log-ratio denoting how the corresponding DNA copy number varies compared to the normal level. The sign of the value represents either a ‘gain’ (amplification) or a ‘loss’ (deletion) of the DNA segment in the corresponding regions on the chromosomes. To represent both DNA amplification events and deletion events at each spot, the log-ratio values of each spot are split into an amplification group and a deletion group at the spot. Hypergraph is a natural model to represent this relation, i.e. we use two kinds of hyperedges to differentiate the amplification and deletion states of CNVs. For example, in Fig 1 we use S1 to S6 to represent individual samples in the arrayCGH data. From the arrayCGH profile on the left of the figure, sample S1, S2 and S3 have ‘gain’ state in spot1 and sample S4 and S6 have ‘loss’ state in spot1. Thus, sample S1, S2 and S3 are covered by an amplification hyperedge denoted by a solid-edge ellipse and sample S4 and S6 are covered by a deletion hyperedge denoted by a dash-edge ellipse in the figure. The ‘basal’ state is regarded as no event at the spot.

Let \( G(V, E, w) \) be a weighted hypergraph to model arrayCGH data: each patient sample is denoted by a vertex \( v \in V \) and each hyperedge denotes one of the two CNV states (gain/loss) of a spot. The incidence matrix \( H \) between \( V \) and \( E \) are defined by the CNV log-ratio values on the samples. Suppose that the log-ratio value of a spot \( i \) of sample \( u \) is \( L(u, i) \). If \( L(u, i) > 0 \) and the state is ‘gain’, \( H_{ui,g} = L(u, i) \) and \( H_{ui,l} = 0 \) where \( i_g \) and \( i_l \) are the indexes of the gain and loss states of spot \( i \) respectively;\( L(u, i) < 0 \) and the state is ‘loss’ in this spot, \( H_{ui,g} = 0 \) and \( H_{ui,l} = -L(u, i) \). Note, in this case \( H \) is a weighted incidence matrix. Given a hypergraph \( G(V, E, w) \), we define a function \( y \) to assign initial labels to \( V \) in the hypergraph. If a vertex \( v \) is in the positive patient group, \( y(v) = +1 \); if it is in the negative patient group, \( y(v) = -1 \); if \( v \) is a test sample, \( y(v) = 0 \). For normalization, in experiments we set \( y(v) = 1/n_1 \) for positive vertex and \( y(v) = -1/n_2 \) for negative vertex, where \( n_1 \) is the number of positive samples in training set and \( n_2 \) is the number of negative samples in training set. To build a predictive model and identify discriminative CNVs, the HyperPrior algorithm will learn a weighting of all the hyperedges in the hypergraph built from arrayCGH data. To avoid overfitting, we introduce a prior on the weights to be learned by the HyperPrior algorithm. In Fig 4 the amplification (deletion) states of adjacent spots are connected in a prior graph as constraints on the weights in the learning regularization framework of HyperPrior. For example, amplification state of spot 1 is connected to that of spot 2, and the one of spot 2 is connected to that of spot 3, etc.

3. HYPERGRAPH-BASED LEARNING

For cancer outcome prediction, our goal is to find the correct labels for the unlabeled vertices of the test samples in the hypergraph. Let \( f \) be the objective function (vector) of labels to be learned. Intuitively, there are two criteria for learning optimal \( f \): 1) we want to assign the same label to vertices that share many incidental hyperedges in common; 2) assignment of the labels should be similar to the initial labeling \( y \). For criteria 1), we define the following cost function,

\[
\Omega(f, w) = \frac{1}{2} \sum_{e \in E} \frac{w(e)}{d(e)} \sum_{u, v \in e} \left( \frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}} \right)^2
\]

If the predicted labels on the vertices are consistent with the incidences with the hyperedges, the value of \( \Omega(f) \) should be minimized. For criteria 2), we directly measure the 2-norm distance between the vectors of the predicted and the original labels as follows,

\[
||f - y||^2 = \sum_{u \in V} (f(u) - y(u))^2
\]

To introduce the spatial relations of spots as prior knowledge into the hypergraph-based learning, we assume that...
adjacent regions on the chromosomes should receive similar weights on their associated hyperedges with the same states. We define a binary indicator $\delta_{i,j}$ to capture the adjacency relations between a pair of hyperedge $e_i$ and $e_j$. The indicator $\delta_{i,j} = 1$ if the two spots associated with $e_i$ and $e_j$ are adjacent on the chromosome and both $e_i$ and $e_j$ are ‘gain’ or ‘loss’ state hyperedges; otherwise 0. To assign weights to hyperedges consistent with such prior knowledge, we define the following cost function over the hyperedge weights,

$$\Psi(w) = \frac{1}{2} \sum_{i,j=1}^{[E]} \delta_{i,j} \left( \frac{w(e_i)}{\sigma(e_i)} - \frac{w(e_j)}{\sigma(e_j)} \right)^2,$$

where $\sigma(e_i) = \sum_{j=1}^{[E]} \delta_{i,j}$, which is the number of hyperedges adjacent to the hyperedge $e_i$. Minimizing $\Psi(w)$ ensures that hyperedges with the same state associated with adjacent regions on chromosomes will get similarly weighted.

After the prior knowledge is introduced, our task is to minimize the sum of the three cost terms, which is

$$\text{minimize}_{f,w} \Omega(f, w) + \mu||f - y||^2 + \rho \Psi(w) \quad (1)$$

subject to

$$w(e) \geq 0 \quad \text{for } \forall e \in E,$$

$$\sum_{e \in E} h(v, e)w(e) = d(v) \quad \text{for } \forall v \in V,$$

where $\mu$ and $\rho$ are positive real numbers. The intuition of adding $\sum_{e \in E} h(v, e)w(e) = d(v)$ as another constraint is to maintain the hypergraph structure. This optimization problem can be solved efficiently with an iterative algorithm introduced by us in [11].

4. EXPERIMENTS

We tested HyperPrior algorithm on an arrayCGH dataset published in [12]. This dataset contains arrayCGH profiles of 57 bladder tumor samples. Following the data preprocess procedures in [10], we removed the probes corresponding to sexual chromosomes and then construct two types of tumor classification problems: by grade (12 tumors of Grade 1 vs. 45 tumors of higher grades) or by stage (16 tumors of Stage Ta vs. 32 tumors of Stage T2+). As a preprocessing step, we performed a $k$-means clustering ($k = 3$) on the log ratio values of all spots and use the resultant three clusters to define the ‘gain’, ‘loss’, ‘basal’ states. This preprocessing step removes some noise in the data and the clustering result is very stable on the one-dimensional data. The standard SVM with linear and RBF kernels, the original hypergraph algorithm [13] and fused SVM [10] were used as baseline algorithms. To assess the classification performance of all methods, we performed a cross-validation by a leave-one-out (LOO) procedure for the two classification problems.

Fig. 3. Enriched biological functions by the 544 genes at highly weighted chromosome regions at the significance level $p$-value < 0.001. The enriched functions are sorted by $p$-values calculated using the right-tailed Fisher Exact Test.

Table I shows the misclassification rates by the best parameters of all the methods. In both classification problems, HyperPrior algorithm achieved the best classification accuracy, which suggests that the additional cost term introduced from the spatial prior does help sample classification.

The weights assigned by HyperPrior algorithm with optimal parameters for grade classification are visualized in Fig 2. The weights of hyperedges for ‘DNA amplification’ and ‘DNA deletion’ are plotted separately. It is evident in the plots that only scarce chromosomal regions are highly weighted by HyperPrior algorithm. We analyzed the biological functions of the genes located in the highly weighted chromosome regions by Gene Ontology (GO) annotations and pathway analysis with Ingenuity. We investigated whether the genes involve over-represented GO categories and biological pathways that are related with bladder cancer. We selected highly weighted chromosome regions having weights higher than 20 by HyperPrior for amplification states and deletion states separately, and find genes in these selected chromosome regions. These filtering process yields 169 genes with ‘amplification state’ and 375 genes with ‘deletion state’. With these genes together as input, Ingenuity identifies 12 enriched functions scoring a $p$-value less than 1.0e-3. The enriched functions include post-translation modification, antigen presentation, and cellular movement shows strong consistency with those identified by [14] [15] [16]. One of the interesting genes in the highly weighted deletion chromosome regions is tumor protein p53 inducible protein 3 (TP53IP3) at chromosome 2. TP53IP3 is a transcriptional target of tumor suppressor protein p53 (TP53) and plays a role in TP53-mediated apoptosis. Recent study showed that there is an association of TP53IP3 promoter polymorphism with higher-grade bladder cancer [17]. Low frequency of alternations at TP53 responsive TP53IP3 gene polymorphism is also known for its association with lung carcinomas and breast cancer risk [18].
<table>
<thead>
<tr>
<th>LOO errors</th>
<th>SVM (linear)</th>
<th>SVM (RBF)</th>
<th>L1-SVM</th>
<th>fused SVM</th>
<th>Hypergraph</th>
<th>HyperPrior</th>
</tr>
</thead>
<tbody>
<tr>
<td>by grade</td>
<td>0.158</td>
<td>0.158</td>
<td>0.211</td>
<td>0.123</td>
<td>0.193</td>
<td>0.105</td>
</tr>
<tr>
<td>by stage</td>
<td>0.188</td>
<td>0.188</td>
<td>0.271</td>
<td>0.146</td>
<td>0.188</td>
<td>0.104</td>
</tr>
</tbody>
</table>

Table 1. This table shows the misclassification rates in the LOO cross-validation on the bladder cancer dataset with the grade labeling and the stage labeling. For SVMs with linear kernel and RBF kernel, parameters $C = \{10^{-5}, 10^{-4}, \ldots, 10^4, 10^5\}$ and $\sigma = \{10^{-5}, 10^{-4}, \ldots, 10^4, 10^5\}$ were tested. For hypergraph (and HyperPrior) algorithm, parameters $\alpha = \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$, and $\rho = \{10^{-3}, 10^{-2}, \ldots, 10^2, 10^3\}$ (for HyperPrior only) were tested.

5. CONCLUSIONS

In this article, we introduced a hypergraph-based method for sample classification and biomarker selection in array-CGH data. In the experiments, our algorithm performs well in classification and identified interesting gene sets that are possibly cancer-relevant. We plan to investigate further the highly weighted chromosome regions and the associated genes for new discoveries.

6. REFERENCES


