Multiview Comodeling to Improve Subtyping and Genetic Association of Complex Diseases

Jiangwen Sun, Jinbo Bi, and Henry R. Kranzler

***Abstract*—Genetic association analysis of complex diseases has been limited by heterogeneity in their clinical manifestations and genetic etiology. Research has made it possible to differentiate homogeneous subtypes of the disease phenotype. Currently, the most sophisticated subtyping methods perform unsupervised cluster analysis using only clinical features of a disorder, resulting in subtypes for which genetic association may be limited. In this study, we seek to derive a novel multiview data analytic method that integrates two views of the data: the clinical features and the genetic markers of the same set of patients. Our method is based on multiobjective programming that is capable of clinically categorizing a disease phenotype so as to discover genetically different subtypes. We optimize two objectives jointly: 1) in cluster analysis, the derived clusters should differ significantly in clinical features; 2) these clusters can be well separated using genetic markers by constructed classifiers. Extensive computational experiments with two substance-use disorders using two populations show that the proposed algorithm is superior to existing subtyping methods.**

***Index Terms*—Classification, cluster analysis, cotraining, genetic association, multiview analysis, phenotypic subtyping.**

I. Introduction

MANY disease traits are a collection of subtypes demonstrating heterogeneity at the molecular and clinical syndrome levels [1]. Categorizing a disease phenotype clinically has been hindered by the inconsistency of subtyping methods and a lack of validation with objective metrics [2]. There is currently no empirically derived, statistically rigorous method to identify and select optimal subtypes of a disease [3]. We propose an approach aimed at finding homogeneous subtypes that can be of use in clinical diagnosis and at the same time be of value in gene finding efforts.

Multivariate cluster analysis has been the most sophisticated method used in subtyping [4]–[7]. Three main steps have been used in prior subtyping studies: 1) collecting both clinical and genetic data for a group of subjects, 2) identifying subgroups by the application of cluster analysis with either k-means, k-medoids, or hierarchical clustering or their combination to clinical features [4]–[6], and 3) conducting linkage or association analysis for the subtypes derived from the sample [7]. Because the creation of subgroups in the second step is independent of the genetic analysis in the third step, the resultant subtypes may be suboptimal and the association analysis may fail.

In a subtyping study, an objective function may be used to evaluate how strongly the subtypes derived from the grouping are associated with a given set of genetic markers, or how well the subtypes can be separated by the genetic markers. Mathematically, given two sets of variables, clinical features Z and genetic markers X from the same sample, the goal is to partition the sample into subgroups based on pairwise similarities between subjects in Z so that the resultant subgroups y can be classified by X. This problem is different from traditional supervised or unsupervised machine learning problems where labels of subjects are either given precisely or not given at all. In our problem, the labels of subjects need to be derived from the clinical features Z so they can be used to train a classifier with the genetic data X.

In the machine learning literature, the most related work might be the set of multiview data analysis methods, cotraining methods [8], and coclustering methods [9] where multiple groups of input variables are collected for the same set of subjects. When only a small portion of the data is labeled, cotraining improves the classification accuracy by enforcing consistency between the classification decisions of the unlabeled data determined by the models learned independently from each of the views. Nevertheless, cotraining methods are not applicable to the subtyping problem because there are no labeled data to start with. Multiview clustering methods seek groupings of subjects that are consistent across different views. These methods treat the data from the two views equally as the input variables. In the subtyping problem, however, the two views have to be treated differently in that one is used to define the subtypes y and the other is used to explain them. For instance, only a sparse set of genetic risk markers are identified to be associated with a subtype but the subtypes may be defined using many clinical features.

The paper is organized as follows. Section II presents the proposed subtyping methodology, based on which a multiobjective program is derived in Section III together with an algorithm to solve it. Computational results on the problems of subtyping opioid dependence and cocaine dependence are examined in Section IV and we conclude in Section V.

Manuscript received March 1, 2013; revised August 1, 2013; accepted September 5, 2013. Date of publication November 5, 2013; date of current version March 3, 2014. This work was supported by the NIH under Grant DA12849, Grants DA12690, Grants DA22288, Grants AA03510, Grants AA11330, and Grants AA13736. *(Corresponding author: J. Bi.)*

J. Sun and J. Bi are with the Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: javon@engr.uconn.edu; jinbo@engr.uconn.edu).

H. R. Kranzler is with the Treatment Research Center, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104 USA, and also with the Philadelphia VAMC, Philadelphia, PA 19104 USA (e-mail: kranzler@mail.med.upenn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2013.2281362

II. Proposed Methodology

We propose a multiobjective optimization framework to solve the subtyping problem. For a set of cluster labels y, each assigned to one subject, we construct a model as a function of a subject's genetic markers X to approximate the subject's label. The model M is built by minimizing a loss function $\ell(y, X|M_\theta)$ where M_θ is a specific inference model, such as the model of support vector machine (SVM), or logistic regression, and θ denotes the set of its parameters. Since the labels y of subjects are not given beforehand, the labels themselves need to be derived. In other words, we optimize the objective as follows:

$$\min_{y,\theta} \ell(y, X|M_\theta) + \lambda R(M_\theta) \tag{1}$$

for the best y and θ where $R(M_\theta)$ defines the regularization term that controls the complexity of the model M, and λ is a tuning factor to balance between ℓ and R. Note that not every possible labeling y of subjects is a feasible solution of problem (1). The search space of y is confined by the similarity measure defined on the features Z.

Suppose that the classification of subjects y is obtained by partitioning subjects based on a similarity measure that is prespecified on Z. The parameters used in the similarity measure often need to be tuned, such as the parameter σ if a Gaussian similarity $\exp(-\|Z_i - Z_j\|^2/\sigma^2)$ is used, where Z_i and Z_j are the two vectors of clinical features for subjects i and j. Choosing different values of σ or other relevant parameters will produce different clusters of the subjects. In general, we expect that the resultant clusters will be well differentiated from each other and that subjects in the same cluster will be closer than those from other clusters in the Z space. Many metrics have been derived in the literature to measure the quality of clusters, such as the Dunn validity index [10], Davies–Bouldin validity index (DBVI) [11], and Silhouette validation [12]. If a metric $\epsilon(y|\sigma, Z)$ is employed to measure the quality of clusters when using a specific value of σ, the metric corresponds to another objective of the subtyping problem. We hence optimize simultaneously two objectives as follows:

$$\min_{y,\theta,\sigma} \begin{cases} \text{Obj}_1: & \epsilon(y|\sigma, Z) \\ \text{Obj}_2: & \ell(y, X|M_\theta) + \lambda R(M_\theta). \end{cases} \tag{2}$$

We assume that $\epsilon(y|\sigma, Z)$ is a metric to be minimized, or otherwise it can be inverted or negated. The two objectives of problem (2) may not be optimized at the same solution. Thus, it formulates a multiobjective optimization problem.

Multiobjective programming (MOP) is a technique that was developed to solve optimization problems with multiple conflicting objectives. Solving a multiobjective program requires the search for Pareto-optimal solutions [13]. Traditional methods convert multiple objectives into a single objective using certain schemes and user-specified parameters. Two simple and widely used methods for such conversions are the weighted-sum method and the constraint method [13]. The weighted sum method transforms two objectives into a single objective by multiplying each objective with a predefined weight and adding them together as follows:

$$\min \;\; c_1 \text{Obj}_1 + c_2 \text{Obj}_2 \tag{3}$$

where the weights c_1 and c_2 are nonnegative and at least one of them is not zero. If the MOP is not convex, the nonconvex frontier of the Pareto-optimal set cannot be obtained by the weighted-sum method. The constraint method reformulates the MOP by keeping one of the objectives and restricting the rest of the objectives within user-specified limits, such as,

$$\min \;\; \text{Obj}_2, \;\; \text{subject to} \;\; \text{Obj}_1 \le \delta. \tag{4}$$

Our MOP-based subtyping framework follows the constraint method, and can be implemented using any proper cluster analysis algorithm to optimize Obj_1, and any suitable classification algorithm to optimize Obj_2. In the following section, we will instantiate this methodology by utilizing a spectral clustering method [14] and the one-norm SVM [15] in the MOP.

III. Multiobjective Optimization Formulation

A spectral clustering method [14] is employed to search for the cluster assignments of subjects by varying the parameter σ in its Gaussian similarity measure. The DBVI [11], measuring how significantly the resultant clusters differ from each other, serves as Obj_1. The one-norm SVM [15] is used to build a classifier, as a function of the genetic variables X, that separates subjects in different clusters. The loss function used in the one-norm SVM serves as Obj_2. Notice that the framework (2) can be realized in conjunction with other choices of clustering and classification methods.

A. Clustering Algorithm

Spectral clustering is a method based on undirected similarity graph $G = (V, E)$ in which each node in V represents a data point (a subject) and each edge in E is weighted by the similarity between the two connected data points. Partitions of data points represented in the similarity graph can be obtained by cutting the graph into unconnected components with the minimum cost. In a balanced cut, the sizes of these unconnected components should be comparable. Two methods have been proposed to achieve this kind of balanced cut, RatioCut [16] and Ncut [17], that minimize the following objectives, respectively:

$$\text{RatioCut}(C_1, \ldots, C_k) := \frac{1}{2}\sum_{i=1}^{k} \frac{A(C_i, \bar{C}_i)}{|C_i|} = \text{Tr}(H^T LH)$$

$$\begin{aligned} \text{Ncut}(C_1, \ldots, C_k) &:= \frac{1}{2}\sum_{i=1}^{k} \frac{A(C_i, \bar{C}_i)}{\text{vol}(C_i)} \\ &= Tr(T^T D^{-1/2} L D^{-1/2} T) \end{aligned} \tag{5}$$

where C_i is one of the identified components (clusters), $|C_i|$ and $\text{vol}(C_i)$ denote the number of nodes and the sum of edge weights in C_i, respectively, and $\bar{C}_i$ consists of the nodes that are not in C_i. The matrix $A = \{a_{ij}\}$ is the adjacency matrix and a_{ij} measures the similarity between the nodes i and j, D is a diagonal matrix whose ith diagonal element $d_{ii} = \sum_{j:j\neq i} a_{ij}$,

L is the graph Laplacian defined by $L = D - A$, $\mathrm{Tr}(\cdot)$ means the trace norm, and both H and T are matrices consisting of indicator vectors as columns defined as follows:

$$H = \left[\frac{1}{\sqrt{|C_1|}}\mathbb{1}_1, \ldots, \frac{1}{\sqrt{|C_k|}}\mathbb{1}_k\right]$$

$$T = D^{1/2}\left[\frac{1}{\sqrt{\mathrm{vol}(C_1)}}\mathbb{1}_1, \ldots, \frac{1}{\sqrt{\mathrm{vol}(C_k)}}\mathbb{1}_k\right] \tag{6}$$

where $\mathbb{1}_i$ is an indicator vector whose entries equal 1 if the corresponding nodes are in C_i, or 0 otherwise. Finding the global optimal solution to either of these two objectives is NP hard [18]. Their relaxed versions have been defined by allowing the indicator vectors in H and T to take real values. It has been shown that the optimal solutions to the relaxed problems of RatioCut and Ncut are the matrices composed by the eigenvectors corresponding to the first k smallest eigenvalues of L and $D^{-1/2}LD^{-1/2}$, respectively [14].

In spectral clustering, the clusters are determined by the adjacency matrix A which is further determined by a prechosen similarity measure. Spectral clustering is sensitive to changes in the similarity measure [14]. In our approach, we search for the most suitable similarity measure, more precisely, the best value of σ in the Gaussian similarity, to optimize Obj_1 and Obj_2.

B. *Objectives in Our Multiobjective Program*

1) First Objective: Spectral clustering requires an adjacency matrix A that encodes the pairwise similarities between subjects and the desired number of clusters k as its inputs, and outputs the clusters C_i of subjects, $i = 1, \ldots, k$. In our approach, we search for the best value of σ in the Gaussian similarity measure to optimize the DBVI [11] that measures the quality of the clusters. DBVI is a measure related to the ratio of within-cluster distance to between-cluster distance. The lower value of DBVI indicates better quality of the clusters. Hence, we minimize the DBVI as follows using Ncut [17] for the best σ:

$$\min_{\sigma} \ \mathrm{DBVI} = \frac{1}{k}\sum_{i=1}^{k}\max_{i \neq j}\frac{\mathrm{Dist}(C_i) + \mathrm{Dist}(C_j)}{\mathrm{Dist}(C_i, C_j)} \tag{7}$$

where $\mathrm{Dist}(C_i)$ is the average distance from each data point in C_i to the cluster center, and $\mathrm{Dist}(C_i, C_j)$ is the distance between the center of C_i and the center of C_j. These distances are calculated in the Z dimension.

2) Second Objective: For each cluster C_i, without loss of generality, we construct a classifier in the linear form of $f(X) = W^T X + b$ to separate the subjects in C_i from the remaining subjects. The model $W_i^T X + b_i$ specific for cluster C_i is obtained by minimizing the regularized empirical error $\ell(y_i, X, W_i) + \lambda R(W_i)$ where we use a binary vector y_i to indicate the cluster membership: $y_i^j = 1$ if subject X_j is in C_i, or otherwise $y_i^j = -1$, $j = 1, \ldots, n$, for all n subjects. We employ the hinge loss commonly used in SVMs, e.g., $\ell(y_i, X, W_i) = \sum_{j=1}^{n}[1 - y_i^j(W_i^T X_j + b_i)]_+$ where $[a]_+ = 0$ if $a < 0$; otherwise, $[a]_+ = a$, and $R(W_i)$ takes a sparse-favoring form in order to select among features, in particular, ℓ_1-norm $\|W_i\|_1 = \sum_d |W_{id}|$. The ℓ_1-norm shrinks the coefficients W of irrelevant variables to zero [15]. Constructing all of the k classifiers together corresponds to minimizing the overall regularized error as follows:

$$\min_{W_i, b_i, i=1,\ldots,k} \sum_{i=1}^{k}\left[\ell(y_i, X, W_i) + \lambda R(W_i)\right] \tag{8}$$

3) Constrained Conversion: Clearly, the first objective is not convex, which leads to a nonconvex multiobjective program. The constraint conversion method is more suitable to find the Pareto-optimal solutions to this problem. As the subtyping problem seeks to obtain clusters that are interpretable in the X dimension (genetic markers), we model the first objective as a constraint. In other words, we search for solutions that minimize the second objective subject to an acceptable quality of clusters in the Z dimension (clinical features). The following problem (9) is the problem we will solve:

$$\begin{aligned} \min_{\substack{\sigma, W_i, b_i \\ i=1,\ldots,k}} \quad & \sum_{i=1}^{k}\left(\sum_{j=1}^{n}[1 - y_i^j(W_i^T X_j + b_i)]_+ + \lambda\|W_i\|_1\right) \\ \text{subject to} \quad & \frac{1}{k}\sum_{i=1}^{k}\max_{i \neq j}\frac{\mathrm{Dist}(C_i) + \mathrm{Dist}(C_j)}{\mathrm{Dist}(C_i, C_j)} \leq \delta \\ & l_\sigma \leq \sigma \leq u_\sigma \end{aligned} \tag{9}$$

where δ, l_σ, and u_σ are tuning parameters to bound σ.

C. *Proposed Algorithm*

Traditional methods for finding the optimal solution to a constrained optimization problem include deterministic approaches such as gradient-based methods, Newton's methods, and nondeterministic approaches such as simulated annealing [19]. To avoid the difficulty of computing derivatives of the objective function, we design an efficient algorithm based on simulated annealing to solve the converted MOP (9) as depicted in Algorithm 1.

In Algorithm 1, the temperature T starts from a high value, and decreases gradually at each iteration. A probability density function defined according to T is used to search for σ_{new}. The first objective is evaluated after the clusters are obtained. If this objective is within the prespecified limit δ, an SVM model is constructed for each cluster, and the second objective is evaluated. The probability of accepting σ_{new} is calculated via the acceptance probability density function discussed in [20] and defined by the objective values h, h_{new} and the temperature T. If this probability is larger than a number randomly drawn from $[0, 1]$, then σ_{new} is accepted; otherwise, the previous value of σ is retained. Readers can consult [20] for more discussions on simulated annealing.

IV. Computational Results

We applied the proposed algorithm to two real-world datasets that were aggregated from genetic studies of opioid dependence (OD) and cocaine dependence (CD) [4]–[7]. We limited the analysis to European Americans to avoid confounding by population

Algorithm 1 Simulated Annealing for MOP (9)

```
Input: $Z$, $X$, $k$, $\delta$, $M_I$
Initialize: $\sigma$, $T$, $h = 0$;
for $t = 0$ to $M_I$ do
    Calculate Temperature $T$;
    Find a neighbor of $\sigma$ to obtain $\sigma_{new}$ based on $T$;
    Construct adjacency matrix $A$ using $Z$ and the Gaussian similarity with $\sigma_{new}$;
    Obtain clusters $C_i, i = 1, \cdots, k$, by running Ncut with $A$ and $k$;
    Calculate $Obj_1$ in (7) and assign its value to $q$;
    if $q \leq \delta$ then
        Compute $W_i$, $b_i$ for each $C_i$ separately by the one-norm SVM;
        Calculate $Obj_2$ in (8) and assign its value to $h_{new}$;
    else
        Continue;
    end if
    if $probability(h, h_{new}, T) > random(0, 1)$ then
        $h = h_{new}$, $\sigma = \sigma_{new}$;
    end if
end for
Output: clusters $C_{i:1,\ldots,k}$, the values of $Obj_1$ and $Obj_2$.
```

differences in allele frequencies and structure. We compared our approach to an existing subtyping method that performed a sequence of two separate steps: spectral clustering and one-norm SVM classification in the same fashion as in [4]. We refer to this as the sequential subtyping method. The two approaches were compared in terms of the separability of their resultant clusters based on genetic markers.

A. Datasets

Subjects were recruited from multiple sites, including Yale University School of Medicine, University of Connecticut Health Center, University of Pennsylvania School of Medicine, McLean Hospital and Medical University of South Carolina. All subjects gave written, informed consent to participate, using procedures approved by the institutional review board at each participating site.

Opioid-use and cocaine-use behaviors were assessed by two separate components dedicated to the diagnosis of OD and CD, respectively, in a computer-assisted interview process, called the semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA) [21]. The SSADDA variables selected by previous OD and CD subtyping studies [4], [6] were used in the present analysis. Multiple correspondence analysis (MCA) [22] was performed to reduce data. The top MCA dimensions that overall explained more than 80% of data variance were used in cluster analysis.

A total of 1350 single nucleotide polymorphisms (SNPs) selected from 130 candidate genes were genotyped for association tests [23]. For each dataset, we performed quality control as follows. SNPs for which data were available for less than 95% of the subjects, or for which the P-value for Hardy–Weinberg equilibrium was less than 10^{-7} were excluded from further analysis. The minor allele frequency (MAF) of each SNP was calculated within each population. SNPs with MAF less than 0.5% in a population were removed from the association tests for the respective population. The remaining missing entries in the SNP data were imputed.

TABLE I
SUMMARY OF THE OD AND CD DATASETS

Dataset	#cases	#controls	#Vars	#MCA Dims	#SNPs
opioid	827	643	69	13	1185
cocaine	1279	187	68	25	1248

For the OD dataset, we treat opioid users as cases and healthy subjects as controls. For the CD dataset, subjects who were diagnosed with cocaine dependence were treated as cases and healthy subjects who had been exposed to illicit drugs were regarded as controls. Table I summarizes the statistics of the two datasets in terms of the numbers of cases, controls, SSADDA variables (Vars), MCA dimensions (Dims), and SNPs used in the subtyping analysis.

B. Experimental Settings

We utilized the CPLEX optimization package to solve the one-norm SVM, and implemented spectral clustering in MATLAB. Adaptive simulated annealing, an open-source variant of simulated annealing, together with its MATLAB gateway (ASAMIN v1.39) was used to search for the value of σ that optimizes the multiobjective program (9). The parameters δ and λ were set to 0.7 and 0.08, respectively. The upper bound of σ and u_σ was set to a number that led to a pairwise similarity value of at least 0.99, and the lower bound of σ, l_σ was set to the value producing a similarity matrix of the median value less than 0.0001. These tuning steps were based on threefold cross validation.

A typical way to choose a value for σ is to use the median value of all entries in the pairwise distance matrix [14]. We fixed σ to the median value in the sequential method. For both the proposed and sequential methods, cluster analysis was only applied to cases. The resultant clusters were characterized based on important clinical features related to drug use and related behaviors. A generalized estimating equation Wald Type 3 χ^2-test was employed to test the significance of the difference between the resultant clusters in these clinical variables with Bonferroni correction for multiple comparisons.

For each obtained cluster, an SVM model was built to separate cases in the cluster labeled as $+1$ from controls labeled as -1. SVM is sensitive to unbalanced data where the size of a sample with one label is significantly larger than that with another label. To address this problem, we duplicated subjects in the smaller group to make the sample size of the two groups comparable. Let a and b be the dominating and minor groups, respectively, n_a and n_b be their sample sizes, and $t = \lfloor n_a/n_b \rfloor$. We first duplicated each subject labeled by b t times, and then randomly selected $n_a - t * n_b$ subjects from the sample pool composed by all subjects with label b. Tenfold cross validation with stratified case-control split was conducted for every cluster, and receiver

TABLE II
CLINICAL OPIOID-RELATED CHARACTERISTICS OF OPIOID USER CLUSTERS [N(%)]

Behaviors	**Cluster 1** 657(79.44)	**Cluster 2** 170(20.56)	χ^2**(df)**	p**-value**
Age of first use [Mean (SD) in year]	21.15(6.59)	21.67(7.71)	0.58	0.45
Used opioids daily or almost daily	653(99.39)	107(62.94)	65.48	5.55×10^{-16}
Injected opioids intravenously	526(80.06)	50(29.41)	134.40	$< 1 \times 10^{-16}$
Stayed high from opioids for a whole day or more	599(91.17)	103(60.59)	78.05	$< 1 \times 10^{-16}$
Strong desire for opioids made it hard to think of anything else	617(93.91)	50(29.41)	245.63	$< 1 \times 10^{-16}$
Opioid use interfered with work, school, or home life	574(87.37)	39(22.94)	201.13	$< 1 \times 10^{-16}$
Family members, friends, doctor, clergy, boss, or people at work or school objected to opioid use	611(93.00)	52(30.59)	187.13	$< 1 \times 10^{-16}$
Been arrested or had trouble with the police because of opioid use	444(67.58)	23(13.53)	114.34	$< 1 \times 10^{-16}$
Give up or greatly reduced important activities due to opioid use	600(91.32)	48(28.24)	212.67	$< 1 \times 10^{-16}$
Ever treated for an opioid-related problem	610(92.85)	37(21.76)	260.89	$< 1 \times 10^{-16}$
Ever attended self-help group for opioid use	505(76.86)	23(13.53)	141.76	$< 1 \times 10^{-16}$

operating characteristic (ROC) curves were obtained using the test data combined from all folds to evaluate the classification performance. We provide the area under the ROC curve (AUC) in our results to compare the two methods. The AUC reflects the cluster separability based on genetic markers.

Moreover, different analytic approaches, such as SVM, or logistic regression, may identify important SNPs of different associative effects. A larger coefficient for an SNP in the SVM models does not necessarily translate into a smaller p-value in logistic regression. We further tested each of the selected SNPs, i.e., those with no zero coefficients in the SVM models, by a separate logistic regression and evaluated their corresponding p-values to determine the significance of the association with the identified subtypes. Here, logistic regression models were obtained in the similar sampling scheme introduced early to balance the data.

C. Opioid-Use Subtypes

We set the desired number of clusters to 2, so that the resultant clusters were sufficiently large and gave adequate statistical power. The optimal value of σ found by our approach was 5.8.

1) Cluster Clinical Characteristics: We characterized the two clusters obtained with $\sigma = 5.8$ based on 11 important clinical variables depicting opioid use and its consequences. Table II shows that the two clusters differ significantly on almost all of these clinical features, except the mean age of first opioid use. Subjects in Cluster 1 have used opioids more heavily than those in Cluster 2. For example, they had heavier daily use and more intravenous injections. The negative consequences of opioid use, such as "interfering with work" and "been arrested" among subjects in Cluster 1 were much more severe than those for subjects in Cluster 2. Thus, Cluster 1 was a heavy opioid user group, whereas Cluster 2 was composed of moderate opioid users.

2) Associated Genetic Markers: Eight SNPs were associated with Cluster 1 at $p < 1 \times 10^{-3}$ as shown in Table III. An SNP (rs915906) was very close to the empirical threshold ($p < 0.05/1154 = 4.34 \times 10^{-5}$) after Bonferroni correction was applied to address the inflation of type I error due to multiple tests. For Cluster 2, SNP rs6957496 on gene *CHRM2* was significant with a p-value close to 10^{-5}, and it remained significant after Bonferroni correction (empirical threshold: $p < 0.05/1154 = 4.34 \times 10^{-5}$). Odds ratios and the genes where the corresponding SNPs are located are also shown in Table III.

TABLE III
RISK FACTORS (SNPS) ASSOCIATED WITH OPIOID-USE SUBTYPES

	SNP	p**-value**	**Odds Ratio**	**Gene**
Cluster 1	rs915906	5.32×10^{-5}	0.6595	*CYP2E1*
	rs10896065	3.32×10^{-4}	2.0537	*FOSL1*
	rs7940700	4.15×10^{-4}	2.2496	*FOSL1*
	rs755203	5.18×10^{-4}	0.7617	*CHRNA4*
	rs2581206	5.56×10^{-4}	0.7594	*SLC6A11*
	rs698	5.59×10^{-4}	0.7615	*ADH1C*
	rs4077851	7.69×10^{-4}	1.5542	*GABRB2*
	rs2515642	8.02×10^{-4}	0.7294	*CYP2E1*
Cluster 2	rs6957496	1.09×10^{-5}	2.25	*CHRM2*

TABLE IV
COMPARISON ON GENETIC SEPARABILITY OF OPIOID USER CLUSTERS

	Optimal $\sigma = 5.8$		$\sigma = 6.07$	
	N(%)	AUC	N(%)	AUC
Cluster 1	657(79.4)	0.59	600(72.6)	0.50
Cluster 2	170(20.6)	0.85	227(27.4)	0.80

3) Comparison: For the sequential method, we followed the standard approach to selecting σ for spectral clustering [14] and computed the median value of the pairwise distances, which was 1.07. When $\sigma = 1.07$, a very unbalanced partition resulted: 826 in one cluster and 1 in the other, which was not of practical value. In order to find a σ value that gives clusters of similar size, we increased the value of σ several times, and each time by 1 until a proper σ was found. The final value was 6.07. Two classifiers were built to separate cases in each subtype from the controls based on genetic data, respectively, for our approach and the sequential method. The AUC values of these classifiers were compared to evaluate the cluster separability in the genetic view as shown in Table IV. Genetic markers had better predictive power for those clusters obtained by the proposed approach than the sequential method with a larger supporting sample size, thus demonstrating the effectiveness of the proposed method.

D. Cocaine-Use Subtypes

Since a large number of cases were available, we set the desired number of clusters to 3. The optimal value of σ found by our approach here was 1.76.

1) Cluster Clinical Characteristics: The three clusters obtained with $\sigma = 1.76$ were characterized in Table V based on 12

TABLE V
CLINICAL COCAINE-RELATED CHARACTERISTICS OF COCAINE USER CLUSTERS [N(%)]

Behaviors	**Cluster 1**	**Cluster 2**	**Cluster 3**	χ^2**(df)**	p**-value**
	340(33.11)	328(31.94)	359(34.96)		
Age of first cocaine use [Mean (SD) in year]	17.61(4.13)	19.53(5.16)	21.28(6.22)	79.50(2)	$< 1 \times 10^{-16}$
Age of onset of heaviest cocaine use [Mean (SD) in year]	25.95(8.09)	25.82(8.12)	29.47(7.70)	44.48(2)	2.19×10^{-10}
Used cocaine daily or almost daily	329(96.76)	251(76.52)	340(94.71)	73.32(2)	1.11×10^{-16}
Injected cocaine intravenously	311(91.47)	132(40.24)	33(9.19)	298.77(2)	$< 1 \times 10^{-16}$
Stayed high from cocaine for a whole day or more	304(89.41)	210(64.02)	327(91.09)	83.49(2)	$< 1 \times 10^{-16}$
Strong desire for cocaine made it hard to think of anything else	308(90.59)	176(53.66)	332(92.48)	162.45(2)	$< 1 \times 10^{-16}$
Cocaine interfered with work, school, or home life	312(91.76)	139(42.38)	311(86.63)	198.06(2)	$< 1 \times 10^{-16}$
Family members, friends, doctor, clergy, boss, or people at work or school objected to cocaine use	310(91.18)	173(52.74)	324(90.25)	159.72(2)	$< 1 \times 10^{-16}$
Been arrested or had trouble with the police because of cocaine use	223(65.59)	69(21.04)	175(48.75)	127.35(2)	$< 1 \times 10^{-16}$
Give up or greatly reduced important activities due to cocaine use	321(94.41)	179(54.57)	340(94.71)	177.31(2)	$< 1 \times 10^{-16}$
Ever treated for an cocaine-related problem	264(77.65)	91(27.74)	249(69.36)	178.74(2)	$< 1 \times 10^{-16}$
Ever attended self-help group for cocaine use	250(73.53)	89(27.13)	227(63.23)	139.27(2)	$< 1 \times 10^{-16}$

TABLE VI
RISK FACTORS (SNPS) ASSOCIATED WITH COCAINE-USE SUBTYPES

	SNP	p**-value**	**Odds Ratio**	**Gene**
Cluster 1	rs3802280	7.98×10^{-4}	1.8265	*OPRK1*
Cluster 3	rs511895	3.03×10^{-4}	0.6456	*CAT*
	rs722651	4.95×10^{-4}	1.5062	*MPDZ*
	rs7940700	5.87×10^{-4}	0.6585	*CAT*
	rs494024	6.22×10^{-4}	0.6602	*CAT*

important features related to cocaine use and its consequences. Table V shows that the three clusters differ significantly on all the 12 clinical features. Both Clusters 1 and 3 were heavy cocaine user groups compared to Cluster 2 as indicated by almost all of the features. For example, 96.76% and 94.71% of the subjects in Clusters 1 and 3, respectively, ever used cocaine daily or almost daily in comparison with only 76.52% of the subjects in Cluster 2. Even though Clusters 1 and 3 were both heavy user groups, they were distinct on several features, especially on the age of onset and on cocaine intravenous injection rates. Subjects in Cluster 1 started the initial and heavy use of cocaine at much younger age than those in Cluster 3. Cluster 1 had a high portion of subjects (91.47%) who had injected cocaine intravenously in contrast to a much lower rate of that (9.19%) in Cluster 3.

2) Associated Genetic Markers: The results from association tests for the three clusters are provided in Table VI, in which only those SNPs with tested p-values less than 1×10^{-3} are shown. SNP rs3802280 on gene *OPRK1* was associated with Cluster 1 at $p < 1 \times 10^{-3}$. Four SNPs were identified to be nominally associated with Cluster 3 at $p < 1 \times 10^{-3}$. None of the SNPs was identified to be associated with Cluster 2 at $p < 1 \times 10^{-3}$.

3) Comparison: For the CD data, the median value of the pairwise distances was 1.45, which was used as the value of σ in the sequential method. We ran spectral clustering based on the similarity matrix and also obtained three clusters. We compared these three clusters against those obtained by our approach in terms of the cluster separability based on genetic data. We built three classifiers, each used to separate subjects in one of the three clusters from the controls. We computed the average and standard deviation of the classifiers' AUC over a tenfold cross validation, respectively, for the proposed and sequential methods. A box plot in Fig. 1 was drawn for each method. As shown in Fig. 1, classifiers trained on the clusters obtained by the proposed method have a slightly better average AUC value (i.e., separability) and significantly smaller error bar than those obtained on the clusters from the sequential method, which implicates that the proposed approach is better in terms of finding genetically separable clinical clusters than the existing approach.

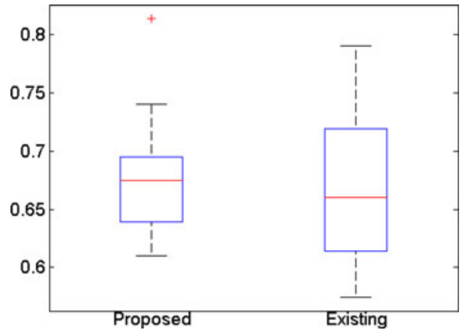

Fig. 1. The comparison of genetic separability of the cocaine user clusters obtained by the proposed method and the sequential method in [4].

V. CONCLUSION

Identifying genes that contribute to risk of complex diseases has been challenging due to two major issues. (1) The diseases have diverse clinical manifestations and complex etiology with both genetic and environmental risk factors. (2) Disease phenotypes are heterogeneous, and homogeneous subtypes have not been optimized empirically. To address these issues, researchers have sought to leverage the technology of cluster analysis to identify clinically homogeneous subtypes that correlate to homogeneous genetic risk factors. Although encouraging results have been obtained, success remains limited because existing methods mismatch the clinical cluster analysis to the goal of genetic association.

We have developed a novel multiobjective programming approach that optimizes two objectives: 1) the cluster-derived subtypes should differ significantly in clinical features; 2) the subtypes can be classified using genetic markers. Our method forms a novel multiview data analytic method that treats the different views differently instead of equally as input views. In our method, the view of clinical features was used to define and derive subtypes of the disease based on cluster analysis, and the view of genetic markers was used to interpret the subtypes based on sparse modeling. Two case studies of subtyping of opioid use and cocaine use, and related behaviors in aggregated samples of European Americans were performed. A comparison between our proposed approach and a typical subtyping method [4] demonstrated the superiority of our approach.

ACKNOWLEDGMENT

The authors would like to thank Dr. J. Gelernter who provided genetic data from NIAAA arrays that enabled the experimental analysis in this study.

REFERENCES

[1] T. Sorlie, “Introducing molecular subtyping of breast cancer into the clinic?” *J. Clin. Oncol.*, vol. 27, no. 8, pp. 1153–1154, 2009.

[2] A. Godfrey, M. Leonard, S. Donnelly, M. Conroy, G. I. ÓLaighin, and D. Meagher, “Validating a new clinical subtyping scheme for delirium with electronic motion analysis,” *Psychiatry Res.*, vol. 178, no. 1, pp. 186–190, 2010.

[3] D. C. Glahn, J. E. Curran, A. M. Winkler, M. A. Carless, J. W. Kent, J. C. Charlesworth, M. P. Johnson, H. H. H. Goring, S. A. Cole, T. D. Dyer, E. K. Moses, R. L. Olvera, P. Kochunov, R. Duggirala, P. T. Fox, L. Almasy, and J. Blangero, “High dimensional endophenotype ranking in the search for major depression risk genes,” *Biol. Psychiatry*, vol. 71, no. 1, pp. 6–14, 2012.

[4] G. Chan, J. Gelernter, D. Oslin, L. Farrer, and H. R. Kranzler, “Empirically derived subtypes of opioid use and related behaviors,” *Addiction*, vol. 106, no. 6, pp. 1146–1154, 2011.

[5] J. Sun, J. Bi, G. Chan, R. F. Anton, D. Oslin, L. Farrer, J. Gelernter, and H. R. Kranzler, “Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors,” *Addict. Behav.*, vol. 37, no. 10, pp. 1138–1144, 2012.

[6] H. R. Kranzler, M. Wilcox, R. D. Weiss, K. Brady, V. Hesselbrock, B. Rounsaville, L. Farrer, and J. Gelernter, “The validity of cocaine dependence subtypes,” *Addict. Behav.*, vol. 33, no. 1, pp. 41–53, 2008.

[7] J. Gelernter, C. Panhuysen, M. Wilcox, V. Hesselbrock, B. Rounsaville, J. Poling, R. Weiss, S. Sonne, H. Zhao, L. Farrer, and H. R. Kranzler, “Genomewide linkage scan for opioid dependence and related traits,” *Amer. J. Human Genet.*, vol. 78, no. 5, pp. 759–769, 2006.

[8] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proc. 11th Annu. Conf. on Comput. Learning Theory*, 1998, pp. 92–100.

[9] A. Kumar and H. D. III, “A co-training approach for multi-view spectral clustering,” in *Proc. 28th Int. Conf. on Mach. Learn.*, 2011, pp. 393–400.

[10] J. C. Dunn, “Well separated clusters and optimal fuzzy-partitions,” *J. Cybern.*, vol. 4, pp. 95–104, 1974.

[11] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[12] P. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, Nov. 1987.

[13] J. Bi, “Multi-objective programming in SVMs,” in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 35–42.

[14] U. von Luxburg, “A tutorial on spectral clustering,” *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[15] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, “1-norm support vector machines,” in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA, USA: MIT Press, 2004, pp. 49–56.

[16] L. Hagen and A. B. Kahng, “New spectral methods for ratio cut partitioning and clustering,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 11, no. 9, pp. 1074–1085, Sep. 1992.

[17] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[18] D. Wagner and F. Wagner, “Between min cut and graph bisection,” in *Mathematical Foundations of Computer Science*, 1993, pp. 744–750.

[19] S. Kirkpatrick, J. C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[20] L. Ingber, “Very fast simulated re-annealing,” *Math. Comput. Modell.*, vol. 12, no. 8, pp. 967–973, 1989.

[21] A. Pierucci-Lagha, J. Gelernter, G. Chan, A. Arias, J. F. Cubells, L. Farrer, and H. R. Kranzler, “Reliability of dsm-iv diagnostic criteria using the semi-structured assessment for drug dependence and alcoholism (SSADDA),” *Drug Alcohol Depend.*, vol. 91, no. 1, pp. 85–90, 2007.

[22] F. Murtagh, “Multiple correspondence analysis and related methods,” *Psychometrika*, vol. 72, no. 2, pp. 275–277, 2007.

[23] C. A. Hodgkinson, Q. Yuan, K. Xu, P. H. Shen, E. Heinz, E. A. Lobos, E. B. Binder, J. Cubells, C. L. Ehlers, J. Gelernter, J. Mann, B. Riley, A. Roy, B. Tabakoff, R. D. Todd, Z. Zhou, and D. Goldman, “Addictions biology: haplotype-based analysis for 130 candidate genes on a single array,” *Alcohol Alcohol.*, vol. 43, no. 5, pp. 505–15, 2008.

Jiangwen Sun received the B.Sc. degree in clinical medicine in 2004 from the Second Military Medical University, China and the M.Sc. degree in computer engineering in 2008 from Nanjing University, China. He is currently a Ph.D. candidate in computer science and engineering at the University of Connecticut, Storrs, CT, USA.

His research interests include health informatics, bioinformatics, data mining, and machine learning.

Jinbo Bi received the Ph.D. degree in mathematics from Rensselaer Polytechnic Institute, USA and the M.Sc. degree in electrical engineering from Beijing Institute of Technology, China.

She is an Associate Professor of computer science and engineering at the University of Connecticut Storrs, CT, USA. Prior to her current appointment, she worked with Siemens Medical Solutions on computer aided diagnosis research. Her research interests include machine learning, data mining, bioinformatics and biomedical informatics.

Henry R. Kranzler received the M.D. degree from Robert Wood Johnson Medical School, NJ, USA, in 1982 and completed a psychiatric residency and fellowship in alcohol research at the University of Connecticut Health Center in 1987.

He is a Professor of psychiatry and Director of the Center for Studies of Addiction at the Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. His research interests include the genetics and pharmacological treatment of alcohol and drug dependence.