A Multi-Objective Program for Quantitative Subtyping of Clinically Relevant Phenotypes

Jiangwen Sun, Jinbo Bi
Department of Computer Science and Engineering
University of Connecticut
Storrs, CT, USA
javon, jinbo@engr.uconn.edu

Henry R. Kranzler
Treatment Research Center
University of Pennsylvania
Philadelphia, PA, USA
kranzler_h@mail.trc.upenn.edu

***Abstract*—Identifying genetic variations that underlie human disease is very important to advance our understanding of the disease's pathophysiology and promote its personalized treatment. However, many disease phenotypes have complex clinical manifestations and a complicated etiology. Gene finding efforts for complex diseases have had limited success to date. Research results suggest that one way to enhance these efforts is to differentiate subtypes of a complex multifactorial disease phenotype. Existing subtyping methods rely on cluster analysis using only clinical features of a disorder without guidance from genetic data, resulting in subtypes for which genotype association may be limited. In this work, we seek to derive a novel computational method based on multi-objective programming that is capable of clinically categorizing a disease phenotype so as to discover genetically different subtypes. Our approach optimizes two objectives: (1) the cluster-derived subtypes should differ significantly on clinical features; (2) these subtypes can be well separated using candidate genes. This work has been motivated by clinical studies of opioid dependence, a serious, prevalent disorder that is heterogeneous phenotypically. Analyses on a sample of 1,470 European American subjects aggregated from multiple genetic studies of opioid dependence show that the proposed algorithm is superior to existing subtyping methods.**

***Keywords*-Subtyping; Cluster analysis; Multi-objective optimization; Gene finding; Opioid dependence**

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.

The proposed method has been applied to a subtyping study of opioid dependence (OD). OD leads to serious medical, legal, social and psychiatric problems. Although the risk of OD is genetically influenced [4], the effort to identify genes and variants that contribute to the risk of OD has limited success because OD is complex in its manifestations, including cognitive, behavioral and physiologic features. The OD phenotype defined by the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) [5] is heterogeneous and may not lend itself readily to gene finding. Decomposition of a complex set of opioid users into homogeneous subgroups can refine the phenotype and enhance genetic analysis [6]. Cluster analysis has been the main method used in subtyping and encouraging results have been obtained [7], [8].

Cluster analysis, however, has been performed only on clinical features without taking into account the rich genetic data that may be available in genetic studies. 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, and (3) conducting linkage or association analysis for the subtypes derived from the sample. 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.

Cluster analysis can create different partitions of the subjects by varying its process parameters. An objective function is often used to measure the validity of partitions or groupings. 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 resulting subgroups y can be classified by X.

The following four sections describe such an approach. Section II describes the proposed subtyping methodology. A multi-objective program is derived in Section III together with an algorithm to solve it. Computational results on the OD subtyping problem are presented in Section IV and

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

conclusions are presented in Section V.

II. Methodology

We propose a multi-objective optimization framework to solve the subtyping problem. In this framework, for a set of 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 optimized. In other words, we optimize the objective in (1).

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

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. Notice 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 pre-specified on Z. The parameters 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 for 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's Validity Index [9] and Davies-Bouldin Validity Index [10]. 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 the following optimization problem (2).

$$\min_{y,\theta,\sigma} \begin{cases} Obj_1: & \epsilon(y|\sigma, Z) \\ Obj_2: & \ell(y, X|M_\theta) + \lambda R(M_\theta) \end{cases} \tag{2}$$

We assume that $\epsilon(y|\sigma, Z)$ is a metric to minimize, or otherwise it can be inverted or negated. The two objectives of Problem (2) may not be optimized simultaneously. Thus, it formulates a multi-objective optimization problem.

Multi-objective programming (MOP) is a technique that was developed to solve optimization problems with multiple conflicting objectives. Solving a multi-objective program requires the search for Pareto-optimal solutions [11]. A feasible point is a Pareto-optimal solution if the point is not dominated by any other point in the feasible set. A solution p_1 is said to dominate another solution p_2 if the solution p_1 is no worse than p_2 in all objectives and the solution p_1 is strictly better than p_2 in at least one objective. Traditional methods convert multiple objectives into a single objective using certain schemes and user-specified parameters. Many studies compare different methods of such conversions, and provide reasons in favor of one conversion over another. Two simple and widely used methods for such conversions are the weighted sum method and the constraint method [11].

The weighted sum method transforms two objectives into a single objective by multiplying each objective with a pre-defined weight and adding them together. If the MOP is not convex, the non-convex parts of the Pareto-optimal set cannot be obtained by the weighted sum method. Hence, the constraint method reformulates the MOP by keeping one of the objectives and restricting the rest of the objectives within user-specified limits. In the next section, we will derive an instantiation of this methodology by utilizing a spectral clustering method [12], the one-norm SVM [13] and the constraint method in MOP.

III. A Multi-objective formulation

To derive a concrete form of our framework, we choose to use a spectral clustering method [14] to search for the cluster assignments of subjects by varying the parameter σ used in its Gaussian similarity measure. The Davies-Bouldin Validity Index [10] is used to measure how significantly the resultant clusters differ from each other, serving as Obj_1. We choose to use the one-norm SVM [13] to fit 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 methods and model fitting methods.

A. 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, $i = 1, \cdots, k$. Given k, the resultant clusters are determined by the adjacency matrix which is further determined by a pre-chosen similarity measure. Spectral clustering is sensitive to changes in the similarity measure [12]. In our approach, we search for the most suitable similarity measure, more specifically, the best value of σ in the Gaussian similarity, to optimize Obj_1 and Obj_2. We use the Davies-Bouldin Validity Index (DBVI) [10] to measure the quality of the clusters. DBVI is a measure related to the ratio of within-cluster distance to between-cluster distance, which can be calculated as follows:

$$DBVI = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \frac{Dist(C_i) + Dist(C_j)}{Dist(C_i, C_j)} \tag{3}$$

where $Dist(C_i)$ is the average distance of data points in C_i to its cluster center, $Dist(C_i, C_j)$ is the distance between the center of C_i and the center of C_j. These distances

are calculated in the Z dimension. The smaller the DBVI, the better the quality of the clusters. Hence, we minimize the DVBI as in (4) using symmetric normalized spectral clustering [14] for the best σ.

$$\min_{\sigma} \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \frac{Dist(C_i) + Dist(C_j)}{Dist(C_i, C_j)} \tag{4}$$

B. Second Objective. Without loss of generality, for each cluster C_i, we construct a classifier in the linear form of $f(X) = W^T X + b$ to separate the subjects in C_i from the rest 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 $y_i^j = -1$, $j = 1, \cdots, n$, for all n subjects. We employ the hinge loss commonly used in SVMs, e.g., $\ell(y_i, X, W_i) = \sum_{j=1}^{n} \left[1 - y_i^j \left(W_i^T X_j + b_i\right)\right]_+$ where $a_+ = 0$ if $a < 0$, otherwise $a_+ = a$, and $R(W_i)$ takes a sparse-favoring form for the purpose of variable selection, such as ℓ_1-norm $||W_i||_1 = \sum_d |W_{id}|$. The ℓ_1-norm shrinks the coefficients W of irrelevant variables to zero [13]. Constructing all of the k classifiers together corresponds to minimizing the overall regularized error as follows:

$$\min_{W_i, b_i, i=1,\cdots,k} \sum_{i=1}^{k} \left[\ell(y_i, X, W_i) + \lambda R(W_i)\right] \tag{5}$$

C. Constrained Conversion. Clearly, the first objective is not convex, which leads to a non-convex multi-objective 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 (6) is the problem we will solve.

$$\begin{aligned} \min_{\substack{\sigma, W_i, b_i \\ i=1,\cdots,k}} \quad & \sum_{i=1}^{k} \left(\sum_{j=1}^{n} \left[1 - y_i^j \left(W_i^T X_j + b_i\right)\right]_+ + \lambda ||W_i||_1 \right) \\ \text{subject to} \quad & \tfrac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \frac{Dist(C_i) + Dist(C_j)}{Dist(C_i, C_j)} \leq \delta \\ & l_\sigma \leq \sigma \leq u_\sigma \end{aligned} \tag{6}$$

where δ, l_σ and u_σ are tuning parameters.

Proposed Algorithm Traditional methods for finding the optimal solution to a constrained optimization problem include deterministic approaches, such as gradient-based method, Newton's method, and non-deterministic approaches such as simulated annealing [15]. To avoid the difficulty of calculating derivatives of the objective function, we design an efficient algorithm based on simulated annealing to solve Problem (6). Algorithm 1 depicts the procedure used to solve the converted MOP (6).

Algorithm 1 Simulated Annealing for MOP (6)

Input: Z, X, k, δ, M_I
Initialize: σ, T, $o = 0$;
for $t = 0$ **to** M_I **do**
  Calculate Temperature T;
  Find a neighborhood of σ, i.e., σ_{new} based on T;
  Construct adjacency matrix A using Z and the Gaussian similarity with σ_{new};
  Obtain clusters $C_i, i = 1, \cdots, k$, by running Spectral Clustering with A and k;
  Calculate Obj_1 in (4) and assign its value to q;
  if $q \leq \delta$ **then**
    Learn W_i, b_i for each C_i separately by the one-norm SVM;
    Calculate Obj_2 in (5) and assign its value to o_{new};
  else
    Continue;
  end if
  if $p(o, o_{new}, T) > random(0, 1)$ **then**
    $o = o_{new}$, $\sigma = \sigma_{new}$;
  end if
end for
Output: clusters $C_{i:1,\ldots,k}$, the values of Obj_1 and Obj_2.

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 pre-specified 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 [16] defined with the objective values o, o_{new} and the temperature T. If this probability is larger than a number randomly drawn from $[0, 1]$, we accept σ_{new}; or otherwise retain the old one. Readers can consult with [16] for more discussions on simulated annealing.

IV. Computational Result

We provide computational results for a sample of subjects recruited for a study of the genetics of OD. We first describe the data and the preprocessing steps used in our experiments. Then we discuss the background of OD subtyping and its challenges, together with our experimental design. Last, we report the clinical characteristics of the clusters that result from our approach, the genetic risk factors identified, and results of comparing the new method against a regular subtyping method.

A. Data & Preprocessing

A total of 1,470 European American subjects were aggregated from multi-site genetic studies of OD and cocaine dependence. All subjects gave written, informed consent to participate, using procedures approved by the institutional review board at each participating site. Subjects were assessed with the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA), a computer-assisted interview instrument [17]. Of the 1,470 subjects, 827 were identified as opioid users and 643 as non-opioid users based on whether they reported having used opioids more than 11 times in their lifetime.

The OD diagnosis section of the SSADDA contains 23 questions, which resulted in 220 variables. These variables provided information about opioid use and related behaviors such as the age of onset of opioid use, frequency of opioid use, and the occurrence of psychosocial and medical consequences of opioid use. Among the 220 variables, 69 were identified previously as key features for the purpose of OD subtyping [8], and were used in the current analysis. The key features included 55 categorical and 14 continuous variables. The 14 continuous variables were discretized into several levels. Multiple Correspondence Analysis (MCA) [18], a dimension reduction method, was used to reduce the 69 variables into 13 principal dimensions. We retained all of the MCA dimensions, which accounted for more than 10% of the data variance.

A total of 1,212 single-nucleotide polymorphism (SNP)s from 130 candidate genes were genotyped for all of the subjects. The 130 candidate genes were selected as candidates for risk of addiction and the related phenotypes of anxiety and depression. Of the 1,212 SNPs, 25 were removed from the analysis because their call rates were less than 95%. Thus 1,185 SNPs were used in our analysis. The remaining missing entries (0.08% of all data entries) in the SNP data were imputed randomly.

B. Settings & Design

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) were used to search for the value of σ that optimizes the multi-objective program (6). Non-opioid users had all negative values for the OD variables. Thus, only opioid users were used in our cluster analysis based on the 69 key variables. Non-opioid users, however, were used in the construction of classifiers with SNPs. We set the desired number of clusters to 2 so that the resultant clusters were sufficiently large and possessed adequate statistical power. The parameters δ, λ were set to 0.7 and 0.08 respectively, the upper bound of σ, u_σ was set to a value that led to a similarity matrix in which entries were at least 0.99, and the lower bound of σ, l_σ was set to the value producing a similarity matrix in which the median was less than 0.0001. These tuning steps were based on 3-fold cross validation.

We built an SVM model to separate each of the resultant clusters. Multiple designs exist to train the classifiers. In one design, each model is trained using all subjects and a label of 1 is assigned to subjects in the cluster and -1 to those outside the cluster. Another design includes subjects in a specific cluster and those who did not use opioids and labels them 1 and -1 respectively. The selection of the design depends on the practical needs of a study. In the present study, because we used a set of control subjects, the set of non-opioid users, we built an SVM classifier to separate subjects in each opioid user cluster from non-opioid users.

SVM is sensitive to unbalanced data, i.e., where the size of a sample with one label is significantly larger than that with another label. To address this problem, we duplicate subjects in the smaller cluster to make the sample size of the two clusters comparable. Let a and b be the dominating and minor clusters, respectively, n_a and n_b be their sample sizes, and $t = \lfloor n_a/n_b \rfloor$. We first duplicate each subject labeled by b t times, and then randomly select $n_a - t * n_b$ subjects from the sample pool composed by all subjects with label b. The optimal value of σ found by our approach was 5.8.

C. Cluster Clinical Characteristics

We characterized the two clusters obtained with $\sigma = 5.8$ based on 11 important opioid use and consequence variables. A generalized estimating equation (GEE) Wald Type 3 χ^2-test was employed to test the significance of the difference between the resultant clusters on these variables with Bonferroni correction for multiple comparisons ($p < 0.05/11 = 0.0045$). The results are included in table IV-C which shows that the two clusters differ significantly on almost all of the important 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 are much more severe than those for subjects in Cluster 2. Thus, Cluster 1 consisted of heavy opioid users, while Cluster 2 was composed of moderate opioid users.

D. Associated Genetic Markers

In total, 333 and 316 SNPs received non-zero coefficients in the one-norm SVM models trained based on subjects in Cluster 1 and Cluster 2 respectively. Larger coefficients in the SVM model do not necessarily lead to a more significant association (smaller p-value) between predictors (SNP) and respondences (phenotype). Thus, we tested each of the selected SNPs with logistic regression and checked their corresponding p-values to determine how significantly they were associated with the identified subtypes. Similar to the

Table I
CLINICAL CHARACTERISTICS OF RESULTANT CLUSTERS [N(%)]

Behaviors	**Cluster1** 657(79.44)	**Cluster2** 170(20.56)	**chi-square**	p**-value**
Mean age of first use 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	$< 1 \times 10^{-16}$

Table II
RISK FACTORS (SNPS) ASSOCIATED WITH CLUSTER1

SNP	p**-value**	**Odds Ratio**	**Gene**
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

Table III
RISK FACTORS (SNPS) ASSOCIATED WITH CLUSTER2

SNP	p**-value**	**Odds Ratio**	**Gene**
rs6957496	1.09×10^{-5}	2.25	CHRM2

SVM models, we trained logistic regression models using subjects in the opioid user cluster versus non-opioid users. Prior to analysis, we removed 32 and 35 SNPs for Cluster 1 and Cluster 2, respectively, because they had a minor allele frequency (MAF) less than 0.5%. We also deleted one SNP for Cluster 1 because it was not in Hardy-Weinberg Equilibrium (i.e., $p < 1 \times 10^{-7}$). Eight SNPs were associated with Cluster 1 at $p < 1 \times 10^{-3}$ as shown in Table II. One 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, one SNP shown in Table III 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 II and table III.

E. Comparison

We compared the proposed method against a subtyping method that first employed cluster analysis, such as spectral clustering, followed by genetic analysis (only after the clusters were generated). Specifically, we compared the clusters resulting from $\sigma = 5.8$ with the result generated by the typical parameter tuning process in spectral clustering [12]. A typical way to choose a value for σ is to use the median value of all entries in the pair-wise distance matrix. Then for each cluster created by the regular method, same as the scheme in our method, an SVM model was built with subjects in the opioid user cluster labeled as 1 and non-opioid users labeled as -1. To ensure a fair comparison, we used the same technology introduced in Section IV-B to deal with the unbalanced data when building the model. The performance of the models resulting from our approach and the comparison method was evaluated by 10-fold cross validation and measured using Receiver Operating Characteristic curves (ROC). We provide the area under the ROC curve (AUC) in our results to compare the methods.

Following the standard approach to selecting σ for spectral clustering [12], we used the median value of the distance matrix computed on our data, which was 1.07. With $\sigma = 1.07$, a very unbalanced partition was created: 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 1.07 several times, and each time by 1 until a proper σ was found. The final value was 6.07. The results were compared as shown in Table IV. Both our method and the regular method created two clusters, a large one and a small one. Models trained with the two large clusters (Cluster 1 for both methods) had comparable performance. However, the Cluster 1 obtained by our method was significantly larger than that created by $\sigma = 6.07$. Superiority of the new method is evident in that it had the same predictive power of the regular method, but with a larger supporting sample size. Models built for Cluster 2 in our method had better separation performance than those trained with the regular method. The comparison implies that the search result of $\sigma = 5.8$ by our algorithm is better than $\sigma = 6.07$, demonstrating the superiority of the proposed approach.

Table IV
COMPARISON ON CLASSIFICATION PERFORMANCE

	$\sigma = 5.8$		$\sigma = 6.07$	
	N(%)	AUC	N(%)	AUC
Cluster1	657(79.4)	0.59	600(72.6)	0.59
Cluster2	170(20.6)	0.85	227(27.4)	0.80

V. CONCLUSION AND DISCUSSION

It has been difficult to identify genes contributing to risk of complex diseases, especially psychiatric disorders, including substance dependence. This failure is due to two major issues: (1) diverse clinical manifestations and complex etiology with both genetic and environmental risk factors; (2) disease phenotypes that 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 homogeneous subtypes that are expected to correlate to homogeneous risk factors. Although encouraging results have been obtained, the success remains limited because existing methods mismatch the clinical cluster analysis to the goal of genetic association. In this paper, we seek to define clinical subtypes with guidance from genetic data by developing a novel multi-objective programming approach that optimizes two objectives: (1) the cluster-derived subtypes should differ significantly on clinical features; (2) the subtypes can be classified using genetic markers.

A case study of subtyping of opioid use and related behaviors in an aggregated sample of 1,470 European Americans was performed and is discussed here. A comparison between our proposed approach and a typical subtyping method demonstrated the superiority of our approach. Two opioid user clusters were obtained from our analysis with sample size of 657 and 170. The two clusters differ significantly on important clinical features as shown in Table I. Moreover, we found significant association of a SNP with the moderate opioid user cluster (Cluster 2) after correction for multiple testing. This finding is consistent with genetic variation that protects against opioid use. For the heavy opioid user cluster (Cluster 2), one SNP approached significance and seven SNPs were nominally significant.

There are limitations to our approach. First, our approach is currently not speed efficient as both simulated annealing and the optimization solver for the one-norm SVM are time consuming, especially when the number of variables is in thousands. Second, although the one-norm SVM eliminated approximately two thirds of irrelevant variables, over 300 variables were retained. Thus, much sparser techniques are required to ensure practicability. Finally, σ is not the only parameter that should be tuned in the process. Other parameters, such as the clinical features used in cluster analysis, can also be optimized by our framework to enrich the search space of potential grouping. This approach has the potential of being applied to diseases other than OD, including substance use disorders (e.g. cocaine dependence) and psychiatric disorders (e.g., major depression).

REFERENCES

[1] T. Sorlie, "Introducing molecular subtyping of breast cancer into the clinic?" *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.*, vol. 27, no. 8, p. 1153, 2009.

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

[3] D. C. Glahn, J. E. Curran, and A. M. Winkler et al, "High dimensional endophenotype ranking in the search for major depression risk genes," *Biological psychiatry*, vol. 71, no. 1, pp. 6–14, 2012.

[4] K. S. Kendler, K. C. Jacobson, C. A. Prescott, and M. C. Neale, "Specificity of genetic and environmental risk factors for use and abuse/dependence of cannabis, cocaine, hallucinogens, sedatives, stimulants, and opiates in male twins," *Am J Psychiatry*, vol. 160, no. 4, pp. 687–95, 2003.

[5] *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR*, 4th ed. Amer Psychiatric Pub, 2000.

[6] 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," *Am J Hum Genet*, vol. 78, no. 5, pp. 759–69, 2006.

[7] 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.

[8] J. Sun, J. Bi, G. Chan, D. Oslin, L. Farrer, J. Gelernter, and H. Kranzler, "Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors," *Addictive Behaviors*, 2012.

[9] J. C. Dunn, "Well separated clusters and optimal fuzzy-partitions," *Journal of Cybernetics*, vol. 4, pp. 95–104, 1974.

[10] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[11] J. Bi, "Multi-objective programming in svms," in *International Conference on Machine Learning*, 2003, pp. 35–42.

[12] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, December 2007.

[13] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," in *Neural Information Processing Systems*. MIT Press, 2003, p. 16.

[14] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*. MIT Press, 2001, pp. 849–856.

[15] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science, Number 4598, 13 May 1983*, vol. 220, 4598, pp. 671–680, 1983.

[16] L. Ingber, "Very fast simulated re-annealing," *Mathematical and Computer Modelling*, vol. 12, no. 8, pp. 967–973, 1989.

[17] 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.

[18] H. Abdi and D. Valentin, "Multiple correspondence analysis," *In: Salkind N., editor. Encyclopedia of Measurement and Statistics*, vol. Thousand Oaks, CA: SAGE;, p. 7, 2007.