1-Norm Support Vector Machine for College Drinking Risk Factor Identification

Michael Zuba[†], Joseph Gilbert[†], Yu Wu[†], Jinbo Bi[†] [*] , Howard Tennen[‡] and Stephen Armeli[§]
[†]Department of Computer Science and Engineering, University of Connecticut,
115 North Eagleville Road, Storrs, CT 06269 USA
[‡]Department of Community Medicine and Healthcare, University of Connecticut Health Center,
263 Farmington Avenue, Farmington, CT 06030 USA
[§]Department of Psychology, Fairleigh Dickinson University, Teaneck, NJ 07666
{michael.zuba, joseph.gilbert, yu.wu, jinbo}@engr.uconn.edu,
tennen@nso1.uchc.edu, armeli@fdu.edu

ABSTRACT

College student alcohol misuse is a major public health concern. According to a national survey, about 44% of students engage in high-risk drinking activities. This paper presents a machine learning approach to a secondary analysis of data collected in a college drinking study at the University of Connecticut Alcohol Research Center sponsored by the National Institute on Alcohol Abuse and Alcoholism. Existing alcohol studies are deductive where data are collected to investigate a psychological/behavioral hypothesis and statistical analysis is applied to the data to confirm the hypothesis. However, the collected data often carries information beyond the original hypothesis. Our approach aims to discover knowledge from multivariate data collected at a major university campus, which may or may not confirm the original hypothesis and lead to potentially new insights. The proposed machine learning approach can effectively identify risk and/or protective factors for high-risk drinking that can be used to help detect and address the early developmental signs of alcohol abuse and dependence within college-aged students. We demonstrate the use of a statistical feature selection method, 1-norm support vector machine (SVM), to help classify college students as either heavy or low-risk drinkers and simultaneously select the risk factors for heavy drinkers. Results of our experiments are evaluated by several psychologists to delineate risk or protective factors and the interaction among these factors for college drinking behaviors.

Categories and Subject Descriptors

H.2.9 [**Information Systems**]: Data Mining; I.2.6 [**Computing Methodologies**]: Learning; I.5.2 [**Computing Methodologies**]: Feature evaluation and selection

General Terms

Algorithm, Performance, Experimentation

Keywords

Machine Learning, Clustering, Support Vector Machine, Feature Selection, College Student Alcohol Consumption

1. INTRODUCTION

The U.S. Department of Health and Human Services (USDHHS) and the U.S. Surgeon General have identified heavy episodic consumption of alcohol in college students as a major public health problem [1]. With reports of binge drinking in college students increasing every year, this once considered "harmless rite of passage" has now been reframed as a top public health problem. It is estimated that roughly 90% of the alcohol consumed by youth under the age of 21 in the United States is in the form of binge drinks [2]. The Federal government has called attention to targeting binge drinking among college students in efforts to reduce the rate of frequent binge drinking episodes [1].

The use and misuse of alcohol on college campuses is not a new or surprising reality. College drinking surveys have existed in the United States for at least 50 years. Existing research indicates that roughly 80% of college students drink alcohol and that at least half of college student drinkers engage in heavy episodic drinking [3]. Studies have supported that college students who participate in excessive alcohol intake are more likely to become involved in risky activities resulting in negative consequences. These consequences range from nonfatal and fatal injuries, alcohol poisoning, blackouts, academic failure, violence (which includes rape and assault), to sexually transmitted diseases. Additionally, students participating in the misuse of alcohol are more inclined to participate in criminal activity that may jeopardize future job prospects [4]. While the consequences mainly affect the students themselves, in some cases they also affect fellow students, roommates, and family members. If any link between alcohol dependence and early adolescent drinking

[*]Correspondence to: Jinbo Bi, Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269. Email: jinbo@engr.uconn.edu

IHI'12, January 28-30, 2012, Miami, Florida, USA.

exists then it is important to be able to identify students who exhibit at risk behavior such that appropriate steps can be taken to safe guard their health and life.

Alcoholism is a medical problem characterized as a chronic, often progressive disease with symptoms that include a strong need to drink despite negative consequences, such as serious job or health problems [3]. Like many other diseases, it has a generally predictable course, recognized symptoms, and is influenced by factors, both genetic and environmental, that are being increasingly well defined [5]. While alcoholism is more commonly associated with the adult population it is important to focus on the early developmental stages of alcoholism. It has been suggested by [6] that research on early drinking origins and adolescent problems are beginning to converge on that of the origins of alcoholism. Therefore, addressing the issues of college drinking problems may yield preventive interventions to reduce the chances of students developing alcohol dependence in their adult years.

We perform a secondary analysis on data of a recently-completed college drinking study sponsored by National Institute of Alcohol Abuse and Alcoholism [7]. We construct quantitative models using advanced machine learning algorithms to map from a variety of personality, motive and academic/social factors to alcohol use severity categories. Machine learning algorithms are usually classified into one of the two categories: supervised learning and unsupervised learning. The distinction between these two categories is inferred from how the learner actually classifies the data. A supervised algorithm is defined as one where the classes are predetermined and class labels are pre-assigned to training examples. This means that the classes are a finite set that were previously defined by a human expert. The machine learner's task is to construct mathematical models that separate examples in different classes. These models are then evaluated on the basis of their predictive capacity in relation to measures of the variance in the data itself. Unsupervised learning algorithms are not provided with these predetermined classifications. A learner must develop these classification labels automatically as part of the algorithm process. This type of learner has to search for similarities between subsets of data in order to determine whether or not they can be characterized as forming a group. For unsupervised learning, these groups are actually named as clusters. One of the major unsupervised learning problems is cluster analysis. Both supervised and unsupervised learning techniques are employed in our analysis to distinguish students drinking behavior and the related risk factors.

In particular, we focus on an advanced statistical feature selection approach, 1-norm support vector machine (SVM), for college drinking risk factor selection. We display the new findings obtained through our analysis procedure which combines 1-norm SVM with cluster analysis. These results present new factors that predict whether a student is a heavy drinker, or a low-risk drinker, and the interaction of these factors may help to identify students who might be at risk for long term development of alcohol dependence in adult life. Our feature selection method is used to identify, among various aspects of the assessment completed in the study, the protective and risk factors that motivate or deter alcohol consumption. Explanations of the selected data mining techniques as well as a discussion of the experimental results are supplied.

The rest of the paper is organized as follows: in Section 2 we provide the background by further discussing the college drinking problem and some successful examples of using SVM in analyzing medical and public health problems. In Section 3 we discuss the analysis method and related machine learning algorithms. In Section 4 we discuss our experimental results and interpret our results. Section 5 presents conclusions and what our results imply for future research.

2. RELATED WORK

College drinking is an important social and medical problem that warrants continued attention. College students live a stimulating and stressful environment. There are many factors that can contribute to student alcohol consumption and abuse, including peer pressure, drinking to enhance stimulation, and drinking to cope as well as academic factors such as overwhelming stress and anxiety, heavy course loads, and the inability to balance tasks. However, it is still unclear as to which factors are influential in producing a heavy drinking student. Understanding what makes a student drink excessively during their college years and what makes another student drink lightly will help to facilitate the development of interventions that target the distinguishing etiological features [8].

Many techniques and study designs exist, both conceptually (hypothesis driven) and empirically, to identify risk and protective factors for risky college student drinking. Statistical methods are widely used to analyze the research data to test research hypothesis and elucidate alcohol-related factors such as drinking motives [7, 9, 10] in addition to assessing the outcomes of various interventions [11, 12]. These analytic techniques range from descriptive statistics (e.g. mean and standard deviation), parametric inference models [13, 14, 15] and factor analysis [9, 16, 17] to structural equation modeling [8]. These techniques focus on modeling performance, i.e., how accurately the model fits the collected data. Instead, our machine learning techniques focus on predictive performance, i.e., how well the resulting model predicts future cases. SVMs have been widely used in other applications and have been proven to be both powerful and accurate in creating predictive models. In Sections 2.1 and 2.2, we describe the alcohol use problem in more detail by introducing two previous studies on alcohol use that are directly related to the setting of our analysis. In Section 2.3 and 2.4, two successful examples of applying SVM to medical problems are discussed. The success in these studies led us to develop our own SVM algorithm adapted to the analysis of college drinking.

2.1 Motivational Studies of College Drinking

A variety of motivational studies have been performed for college drinking [18, 19, 20]. A typological approach is presented in [8] to identify patterns of alcohol consumption in college freshmen. The study examined quantity and frequency of consumption and alcohol-related problems. Important differences such as heavy alcohol consumption on few occasions versus many frequent occasions can hide potential drinking patterns. The authors examined the number of potential problems in one's life and its correlation to drinking patterns. Separate analysis was conducted for each gender.

A total of 530 freshman students (36% men and 64% women) were recruited through a Psychology Department at a public university. The study was started 5 weeks into the fall semester at which point 62.3% of women and 65.8% of men indicated drinking at least one drink a month. Each student participated in a group-testing format in which they reported their typical alcohol use, alcohol related problems, and reasons for drinking.

Two main patterns emerged from their analysis which can be considered as typical patterns for college student alcohol consumption. The first pattern was that light drinkers or abstainers reported relatively few problems in their life. The second pattern found was that a larger group formed and was associated with moderate to high quantities and frequencies of alcohol consumption as well as a moderate number of problems in their life. The authors also discovered that drinking motives such as drinking to enhance, or drinking to cope can contribute to the development of a drinking model that classifies students. Furthermore, there exist some minor differences in drinking between genders. Similar patterns are also found in our analysis with statistical cluster analysis.

2.2 Factors for Abuse Prevention

Risk and protective factors are reviewed in [21] for alcohol and drug abuse problems in adolescence and early adulthood. Their method focuses on a risk-focused approach that requires the identification of risk and protective factors in order to develop effective substance abuse programs and recommendations for future research and practice. The focus of their paper is on reviewing existing knowledge and methods.

The authors state that it is difficult to ascertain which risk factors or interaction of risk factors are most influential to substance abuse [21]. It is therefore hard to provide an appropriate plan for prevention. They proposed that risk factors can be divided into two categories. The first is a societal and cultural category which provides the normal expectations for behavior. The second group is individual and interpersonal environments which provide factors relative to family, school, and peer life.

The authors conclude that a risk-focused approach to alcohol and drug abuse prevent holds promise for identifying effective prevention strategies [21]. Our work represents a preliminary effort to quantify the interplay of the significant risk factors identified in a motivational study setting with a potential to generalize to unseen cases.

2.3 Cancer Gene Selection using SVM

An effective method of gene selection utilizing statistical feature selection and SVM methods is proposed in [22]. In this paper, authors attempt to tackle the problem of selecting a smaller subset of genes from a broad pattern of gene expression data that was recorded on DNA micro-arrays. The authors present a classifier that is capable of being used for genetic diagnosis, specifically colon cancer, and for drug discovery. Although this work is not directly related to the college drinking problem it is important to review to provide an understanding of how powerful a SVM-based feature selection scheme can be in medical scenarios for knowledge discovery.

In this work the authors use a technique known as Recursive Feature Elimination (RFE) which is a type of feature selection method. The goal of RFE is to train a classifier that optimizes its weights in regards to the cost function of the linear SVM. It then computes the ranking criterion for all the features and removes the features with the lowest ranking scores [22]. In order to test the idea of using the weights of a classifier to produce a feature ranking scheme, the authors use a linear SVM, called SVM-train. This method is then tested on two different cancer databases to find the top ranked genes in each database. These top ranked genes discovered by the SVM were then verified to all have a plausible relation to cancer [22]. The authors also have concluded that their results are more robust to data overfitting than many other methods such as combinatorial search.

2.4 Genetics of Alcoholism using SVM

Support vector machines have also been used [23] to analyze a microsatellite marker dataset from an alcoholism genetics study in order to classify phenotype variables in divided genomic regions. The authors chose 315 microsatellite markers on 22 autosomal chromosomes with 12 different phenotype variables. These 22 sub-datasets of each chromosome are then merged into a single genome dataset, applied to a chosen phenotype and run in an SVM. This work makes use of a program called mySVM [24] which is a specific implementation of an early SVM introduced by V. Vapnik in 1998 [25]. The results of the SVM were confirmed to be of high level of correctness for each prediction, specifically the authors found 96% correctness in the 4-fold cross validations.

This work proves that the SVM methodology is an effective approach for association studies, data reduction, and for the detection of causal genes in future genetic studies. The authors also suggest that SVMs need to be tailored to the specific application and dataset being used in order to fully maximize their efficiency. This is a direct result of using a pre-implemented SVM machine that was not designed to consider application specific details. Regardless of this fact, the use of SVMs is shown to be both powerful and accurate.

3. METHOD

In this section we will present the details regarding the alcohol survey, the data acquired from the survey, the proposed machine learning approach and our analysis procedure.

3.1 College Drinking Data

The data used in our analysis was collected from a study, completed over a time period of one year, at The University of Connecticut Alcohol Research Center. The subject pool consisted of college-aged students enrolled in the Introductory Psychology course at the University of Connecticut who had reported drinking alcohol at least twice in the past month. A survey instrument was designed [7] and was completed by 530 college students in which 52% were female. Each participant was asked to complete a survey questionnaire approximately one month after starting their school semester. The survey was completed with a distribution of 61% in the fall semester and 39% in the spring semester.

The survey questionnaire consisted of over 100 items to measure various risk factors including drinking motives, academic performance, personality, recent depression and anxiety symptoms, negative life events and some demographic items together with drinking behavior of the last month. With the exception of demographic questions and some school status questions, responses to multiple questions in a group were compacted into a risk factor variable with rating scores. Thirteen such composite factor variables were acquired as listed in Table 1 from Factor 9 to Factor 21. Most of the original questions in the survey take ratings from 1 to 7 scales with 1 the lowest level and 7 the highest level. In our analysis, for instance, Factor 15, Beck Depression Inventory (brief version) was calculated by averaging the rating scores for 13 questions asked related to depression, leading to a numerical value ranging from 1 to 7.

Table 1: Risk Factors

Factor ID	Question/Factor
1	School Semester
2	Age
3	Current School Year
4	Dating Status
5	Fraternity/Sorority Status
6	GPA
7	Race/Nationality
8	Religious Preference
9	Drinking to Cope
10	Drinking to Socialize
11	Drinking to Enhance
12	Drinking to Conform
13	Sensation Seeking
14	Neuroticism
15	Beck Depression Inventory
16	Trait Anxiety Inventory
17	Social Anxiety Question
18	Negative Life Events in Last Year
19	Antisocial Personality Traits/ Conduct Disorder
20	Family Social Support
21	Friend Social Support

The variables in our data can be categorized into one of three types: nominal, numerical, or binary. Besides the composite variables which take rating values, there are 8 other risk factor variables listed in Table 1 from Factor 1 to Factor 8. Among them, Age of Student, Current Year in School and Grade Point Average take numerical values. A nominal variable is where all of the attribute values are included in a pre-determined set of labels and do not imply any measurement or ordinal relation. Nominal attributes in our data are Factors 7 and 8, Race/Nationality and Religious Preference. In order to make this data useable in our machine learning algorithm we needed to create a scheme in which nominal and binary variables can be used in combination with numerical variables.

Besides the various risk factors, the survey also measured the recent drinking behavior. The following questions were asked in the survey: the number of occasions in the past 30 days you have consumed an alcoholic drink; the number of average drinks when you drink in the past 30 days; and the number of occasions you were drunk in the past 30 days. As shown in Tables 2, 3, and 4 a predetermined range was selected and then translated over into unique numerical ratings. Another drinking variable, Alcohol Dependence Symptoms, is a numerical average of 17 rating questions in the survey including questions like “How often have you experienced blackouts (loss of memory for drinking episodes)?” and “How often have you consumed alcohol instead of eating a meal?” Students who are likely to develop alcohol dependence symptoms will have a higher value such as 3 (the highest is 5) and students who have little to no symptoms will have a lower value such as 0 or 1.

Table 2: Number of drinking occasions in the past 30 days student consumed alcohol

Drink Occasion Range	Unique Ratings
0	0
1-2	1
3-5	2
6-9	3
10-19	4
20-39	5
> 40	6

Table 3: Average number of drinks in the past 30 days per occasion

Drinks Per Occasion Range	Unique Ratings
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
> 9	9

Table 4: Number of occasions in the past 30 days student was drunk

Occasion of Drunk Range	Unique Ratings
Never	0
1-2	1
3-5	2
6-9	3
10-19	4
20-39	5
> 40	6
No drinks in the past 30 days	7

3.2 Machine Learning Analysis

Existing statistical analysis techniques for risk identification use the entire collected data set to build a model in order specify statistically significant factors. The analysis result often fails to adapt to other studies that also collect the same measurements at a different scenario. In our analysis, we emphasize the generalization performance of a classification model, i.e., the prediction accuracy on unseen cases. The overview of our analysis is depicted in Figure 1. Our analysis approach is data-driven and based on the use of a support vector machine [25]. A K-Medoids cluster analysis [26] was first applied to the 4 drinking behavior variables, shown in Table 5, to classify each student as one of the three types, 3 - a heavy drinker, 2 - a moderate drinker, or 1 - a low-risk drinker. Then each student in the data set will be labeled accordingly with a drinking type label from 1 to 3. We then build 3 classifiers, each used to separate one drinking type from the rest. The labeled data is partitioned by gender and then each group is split into a training set containing 2/3 of the data of the group and a test set containing the other 1/3 of the data. This scheme allows for each classifier to be tested on a data set independent of the data set used when training the classifier.

The 1-norm SVM approach [27] has a tuning parameter c which is chosen according to cross validation performance on training data. A linear model is then constructed by SVM using the chosen c value. The resultant classifier is evaluated on the test data to report classification accuracy measured by the Receiver Operating Characteristic (ROC) curves [28]. This process is repeated n times by randomly partitioning the data n times. The n resulted linear models can be aggregated into a bagging classifier [29]. Bagging has been proved to reduce the influence of sample variance on the construction of an accurate model. Figure 1 provides a flow chart of our scheme. As SVMs by design are binary classifiers, which separates items between two classes, the training process depicted in Figure 1 is repeated three times per gender to construct three distinctive linear models, one for separating one of the three classes from the rest. In further subsections we expand upon our experimental process and techniques used.

3.2.1 *Data Preprocessing*

The analysis process for any data mining technique often requires pre-processing of the data in order to properly analyze the data by machine learning algorithms. In our data, each row represents a unique participant or a single completed survey and each column entry represents a participant's rating value to a risk factor or a drinking behavior variable. The data consists of several attribute types. We are able to employ all the variables of different attribute types by applying Multiple Correspondence Analysis (MCA) [30], which is the generalization of Principle Component Analysis (PCA) when variables to be analyzed are categorical instead of numerical, to analyze the categorical variables. The MCA technique operates on an indicator matrix $Znxk$ where entries are either 1 or 0, n is the number of samples, and if each variable has k_j categories, k is the sum of k_j over all variables. In general, MCA provides a geometric model of the data and summarizes the relations between the categorized variables [30]. This step of pre-processing is more sophisticated than the typical schemes used in the analysis of alcohol-use survey data that simply assign a numerical number to each category of a categorical variable.

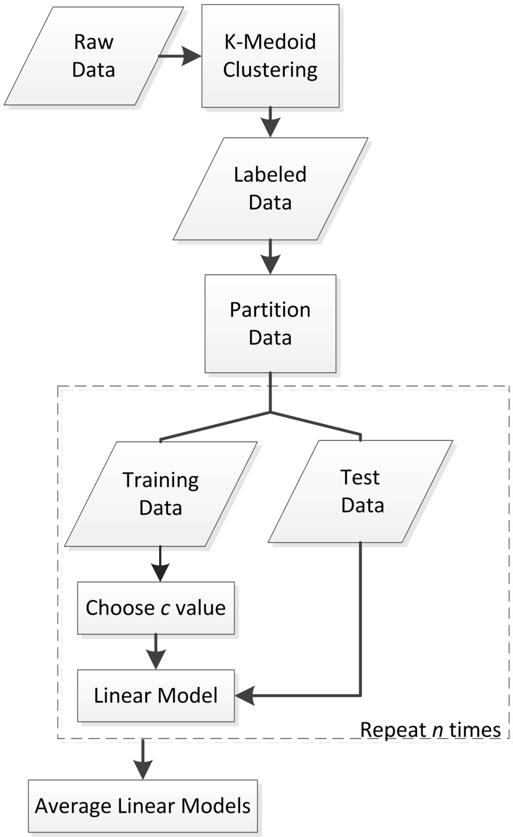

Figure 1: Flowchart of the experimental design

Additionally, we had to deal with the issue of noisy data. Noisy data can be the result of various issues such as when a participant answers a question with an invalid response or the data being wrongly input when entering the results into a spreadsheet or database. For this reason, we decided to remove any student from the survey that had any missing responses or invalid responses to any questions on the survey. After this step, 513 students remained in our analysis. Some questions had data that was represented by a 0 or 1 while other questions had data represented as a rank such as 7, or 5. This could create problems in which some factors would be weighted higher than others because they were represented by a higher numerical value. In order to address this issue we preformed a standard normalization for each column corresponding to a risk feature. Each feature is hence normalized to have a mean value of 0 and a standard deviation of 1.

3.2.2 *Cluster Analysis*

A K-Medoids clustering algorithm [26] is applied to the four drinking variables listed in Table 5. The first three drinking variables are shown in Tables 2, 3 and 4 which are presented earlier in Section 3.1. These four features are first normalized using standard normalization and then the K-Medoids algorithm is applied with the number of clusters k set to 3. Figure 2 shows the mean values of the four variables for each of the 3 clusters. Figure 3 shows the median values of the four variables, of each cluster, which are the cluster medoids that the K-Medoids method uses to represent individual clusters. Based on the characteristics shown in the two figures, the identified clusters are well separated in terms of the characteristics of drinking behaviors, and we name the three clusters by heavy drinker, moderate drinker, and low-risk (light) drinker groups.

Table 5: Drinking Variables

Unique Name	Question/Factor
DrinkFreq	Number of occasions in the past 30 days student has drank
DrinkAmnt	Average number of drinks in the past 30 days per occasion
DrunkFreq	Number of occasions in the past 30 days student was drunk
AlcDep	Alcohol Dependence Symptoms

The K-Medoids algorithm performs cluster analysis that is similar to the commonly used K-Means. K-Medoids is used to divide data into disjoint clusters. This algorithm attempts to minimize a squared error averaged over all clusters. The square error is proportional to the overall distance between the examples in the cluster and an example that is the center or medoid of the cluster [26]. A medoid is considered to be an object of the cluster whose average dissimilarity to all the objects in the cluster is minimal [26]. The difference between K-Medoids and K-Means lies in the choice of data examples chosen as centers and K-Medoids is more robust to noise and outliers. It is considered to be an effective approach to solving any clustering problem.

3.3 Feature Selection

Feature selection [31, 32, 27] is an active research area in data mining communities. The central idea of feature selection is to construct and select a subset of input variables, or features, by eliminating features with little or no predictive information in order to build a good predictive model [33]. Feature selection allows researchers to significantly improve the comprehensibility of the resulting classifier models and build a model that better generalizes to unseen examples. It can be used to facilitate data visualization and data understanding, reduce the measurement and storage requirements, reduce training and utilization times, and can help to overcome the curse of dimensionality to improve predication performance [31].

Feature selection methods are often divided into two types: filter and wrapper methods [34]. The filter approach of selecting variables serves as a preprocessing step to the induction. The main disadvantage of the filter approach is that it completely ignores the effects of the selected variable subset on the performance of the classification algorithm. The wrapper method searches through the space of variable subsets using the estimated accuracy from a classification algorithm as the measure of "goodness" for a particular variable subset. Therefore, the variable selection is being "wrapped around" a particular induction algorithm. These methods have brought some success with induction tasks, but they can be computationally expensive for tasks with a large number of variables. We generally split data into training and validation sets and evaluate the constructed classifiers by the significance of differences in validation errors [14]. The 1-Norm SVM is a type of wrapper method which selects a subset of features simultaneously when building a classifier to achieve the best validation performance. In other words, the features used in the best classifier are the selected ones.

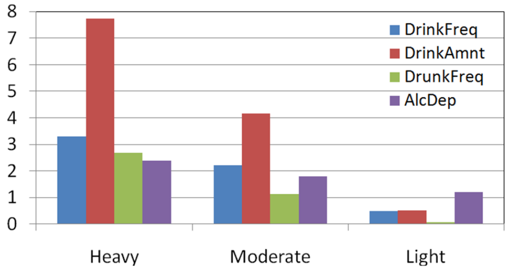

Figure 2: Mean values for each feature in Table 5

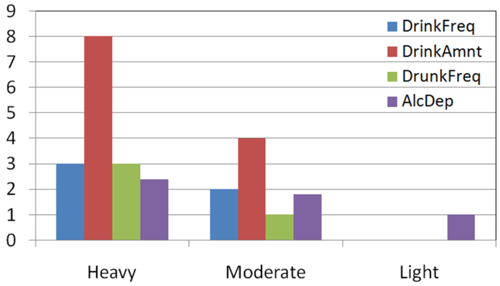

Figure 3: Center values for each feature in Table 5

3.3.1 1-Norm Support Vector Machine

SVM is a supervised learning method, which has the ability to weigh input features according to their relevance to the classification target as determined through the learning process [27]. The SVM classification, in our implementation, constructs a non-probabilistic binary linear classifier. A SVM algorithm can be given a set of data, referred to as the training examples, where every unique entry is marked as belonging to one of two categories or classes. The SVM classification algorithm then builds a model that predicts whether new entries will fall into one category or the other. Most linear SVMs, including the one presented in this paper, often use a constant referred to as the regularization parameter and it is represented by the letter c. This parameter, c, balances the relative influence between the training error (the second term in Eq. (1)) and the model complexity (the first term in Eq. (1)). It controls the trade-off between allowing training errors and forcing rigid margins and therefore SVMs require a search for the optimal value of c.

The 1-norm SVM constructs a classifier based on a linear function of the form of $W^TX+\ b$ where W is the weight vector and X is the input vector representing one student's record by minimizing the following regularized risk function:

$$\sum_{i=1}^{d} |W_i| + c\sum_{j=i}^{N} \epsilon_j \tag{1}$$

where d represents the number of variables in the dataset and N represents the number of students. A sparse SVM simply means that the optimal solution of W is usually constructed based on fewer variables than in classic SVMs. However, the above objective function is not in a canonical form of an optimization problem given the absolute value appears in it. Hence we use a variable transformation to convert the problem into the following equivalent optimization problem:

Minimize:

$$\sum_{i=1}^{d} V_i + c \sum_{j=i}^{N} \epsilon_j$$

Subject to:

$$Y_i(W^T X_i + b) \geq 1 - \epsilon_i, \text{where } \epsilon_i \geq o$$

$$-V_i \leq W_i \leq V_i, \text{where } V_i \geq 0$$

where:

$$|W_i| \leq V_i \rightarrow -V_i \leq W_i \leq V_i$$

By the change of variables from W to V in the objective function that is to be minimized, we impose an upper bound on the magnitude of the weights W associated with each feature. It can be proved that when an optimal solution is found for the problem, the optimal value of V is exactly equal to the absolute value of W. Minimizing the 1-norm penalty in the objective is known to create a sparse weight vector W. In other words, a large portion of weights in W will be driven to 0 at optimality. Hence, only those features that receive non-zero weights in the linear model $W^T X + b$ are selected. In the optimization problem, Y represents the class of each entry and X is the vector of feature values. The variable b represents the offset of the linear model on the Y-axis. The vector X is then taken and used to dot product itself with W, we then add b and get the prediction Y. Thresholding on the predicted values of Y yields the class labels for each student. For example, if Y is greater than or equal to a discrimination threshold, then the specific student belongs to the "positive" class; or otherwise belongs to the "negative" class.

4. EXPERIMENTAL RESULTS

Given our classification approach, we will have three different classes. These classes will be heavy drinkers, moderate drinkers, and low-risk drinkers. SVMs are generally a non-probabilistic binary linear classifier and since we have three different class labels, we need to create a model for each class label. In order to do this we have three duplicate sets of data with different labeling. In each set we pick a class as the main label and combine the other two classes into one label. An example would be labeling the data to construct the heavy drinker classifier. This model is designed to identify students as heavy drinkers and identify features that contribute to the heavy drinking behavior and features that protect people from drinking heavily. We will assign a label of 1 to all students in the heavy drinker class, and 0 to all other students in the other two classes. We perform this kind of combination two more times in such that we will have a total of three models, one for each class. The models are built using the 21 composite risk factors given in Table 1. The raw data was partitioned by gender and therefore we have three different classifiers for each gender. In total we have 6 different classifiers.

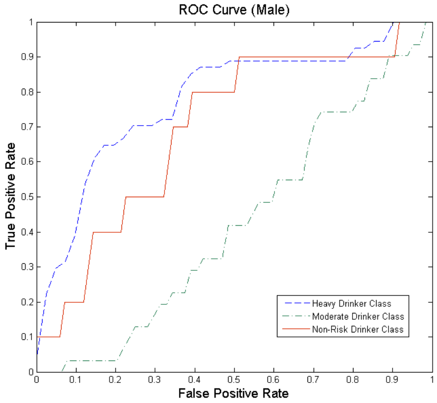

Figure 4: ROC Curves for Male Drinking Classes

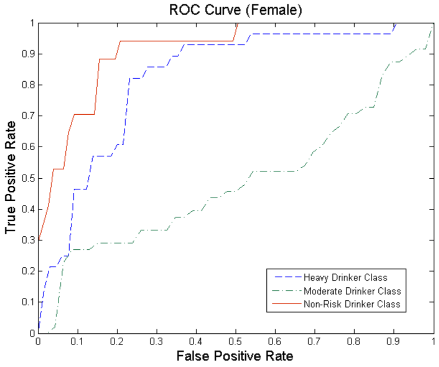

Figure 5: ROC Curves for Female Drinking Classes

4.1 Generalizability

The generalization performance is the main concern of machine learning research. Therefore, in order to validate the accuracy of our models and determine whether or not the weights constructed are important we used a technique known as a Receiver Operating Characteristic (ROC) curve. An ROC curve is a graphical representation or plot of the true positive rate (sensitivity) versus the false positive rate (1-specificity) for a binary classifier system as its discrimination threshold is varied [28]. In a binary classification problem an example is labeled either in "positive" or "negative" class. This results in a total of four possible outcomes. The first case is where the person is predicted to be positive and is in fact positive, known as a true positive. The second case is where the person is predicted to be positive and is in fact negative, known as a false positive. The last two cases are the inverse of the first two; specifically, the person is predicted to be negative and is or is not actually the negative case. These last two cases are known as true negatives or false negatives.

In general, a model with good performance should have a curve that is more towards the top left-hand corner as this implies higher true positive rates at lower false positive rates.

An ROC curve with a straight diagonal line means that the prediction of the classifier is random, and therefore not accurate. In Figure 4 and Figure 5 we show the results of our ROC curves for males and females respectively. We observe that the heavy drinking classifier and low-risk drinking classifier for female performed the best among all models. The heavy drinking and low-risk drinking male classifiers were moderately accurate. The moderate drinker classifiers for both male and female performed poorly and were closer to a random-guessing classifier. However, this is expected as there is a lack of distinguishing features that could appropriately differentiate between the middle class - moderate drinkers and the extreme cases at the two opposite ends: heavy drinkers and light drinkers.

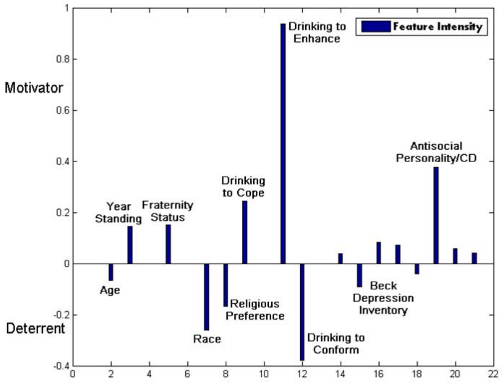

Figure 6: Weight Values for various factors in the Heavy Drinker class (Male)

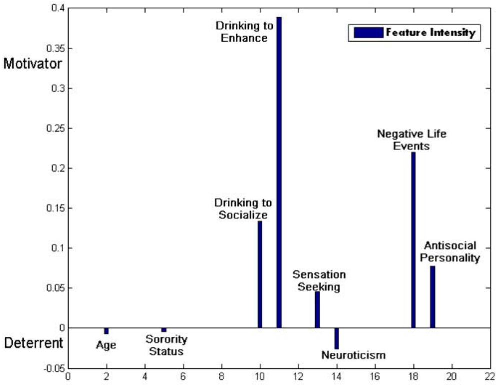

Figure 7: Weight Values for various factors in the Heavy Drinker classifier (Female)

4.2 Identified factors

The SVM training process yields a linear classifier with weights associated with each of the features or the risk factors. These weights specify whether a specific factor is a motivator for drinking heavily (positive weight), whether a factor has no effect (a weight value of 0), or whether the factor is a deterrent for preventing from drinking too much (negative weight). Since our ROC curves show that the moderate drinking classifiers perform poorly, we have omitted further analysis of the weight values used in those classifiers.

Figure 6 and Figure 7 provide the bar plots to show the factors that influence male and female students respectively to become heavy drinkers. Factors that are positive are motivators to drink heavier while factors with negative weight values are ones that deter students from moving towards to the heavy drinker class.

Figure 8 and Figure 9 show the factors that influence male and female students respectively for being resilient to drinking as low-risk drinkers never drink or drink very little. Factors that are positive are motivators for students to drink more. In other words, these factors make students move towards moderate or heavy drinkers. Negative valued factors deter students from drinking, thus maintaining students stay in the low-risk drinker class.

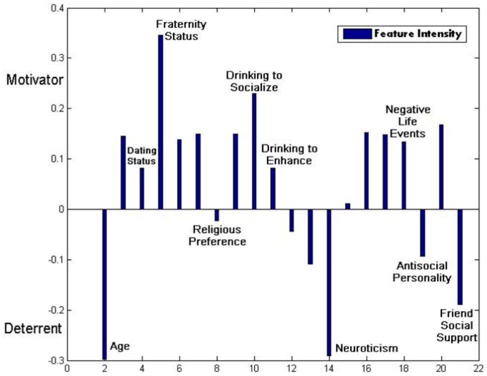

Figure 8: Weight Values for various factors in the Low-Risk classifier (Male)

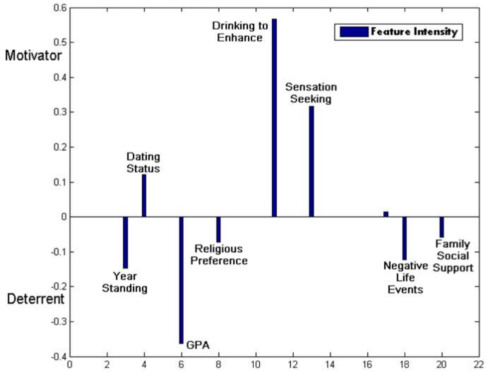

Figure 9: Weight Values for various factors in the Low-Risk classifier (Female)

In Table 6 we have summarized the top identified risk factors that are attributable to "being a heavy drinker" for both male and female. Two risk factors "drinking to enhance" and "antisocial personality" are common between the genders although their influence intensity are different depending on the gender. Elevated negative life events affect female students significantly but impose negligible effects on male students.

Table 7 summarizes the top identified protective factors for being a low-risk drinker for both male and female. One protective factor common between the genders is Age or Year Standing which are two highly correlated features. It shows the older students tend to regulate themselve to drink less among low-risk drinkers. Female students with higher GPA seemed to drink less. Of low-risk drinkers or non-alcohol users in this population, at elevated negative life events, male students tend to drink more but female students tend to withstand drinking although in a very small magnitude.

Table 6: Identified Risk Factors for Heavy Drinkers Ordered by the Descending Order of Their Weights

Gender	Risk Factor
Male, Female	Drinking to Enhance
Male, Female	Antisocial Personality Traits/ Conduct Disorder
Male	Drinking to Cope
Female	Negative Life Events
Female	Drinking to Socialize
Female	Sensation Seeking
Male	Fraternity Status

Table 7: Identified Protective Factors for Low-Risk Drinkers Ordered by the Ascending Order of Their Weights

Gender	Risk Factor
Female	GPA
Male	Neuroticism
Male	Friend Social Support
Male, Female	Year Standing/Age
Male	Sensation Seeking
Male	Antisocial Personality Traits/ Conduct Disorder

4.3 Discussion

Overall, our models produced reasonably good test performance. We can attribute some of the degradation in performance due to one major factor: availability of the data. The sample size of 530, although a common and well-structured size for social and behavioral studies, is statistically still small. After splitting the data by gender we were left with two smaller data sets to analyze. Our results can be validated further if they can be duplicated in other studies. Our models achieved a minimal error rate, which implies an accurate predication rate for future data sets. We can compare our results with that of those from the original study in [7]. The original study found that increased anxiety or depression symptoms affect students who are high in coping motives but low in social enhancement motives by potentially intensifying social vulnerabilities. This would make the student less likely to attend social gatherings in which drinking might occur. Our models appear to correlate with these findings in [7]. Additionally, the original study found that individuals with stronger coping motives would decrease their drinking frequency during periods of elevated negative effect. This correlates with our low-risk drinker class for females but not for males. Moreover, in our findings, antisocial personality traits are an active drive for heavy drinkers to drink more but, surprisingly, a resilience factor for low-risk drinkers to drink less among male students. As the original study noted, the effects of stress and negative affect on college students' alcohol use are complex. Each student falls into various levels of classification based on their coping abilities, social vulnerabilities and appetite for sensation or enhancement affects.

Several of the risk factors we identified are consistent with published results, such as "drinking to enhance." Other factors we identified are not yet well studied in the literature and suggest that more thorough studies are needed to investigate these factors. The emphasis of our study is the generalizability of predictive models to cases that are independent from the data used to obtain the models. Therefore, our SVM models are more likely than any previous models to extend to unseen cases or new studies. Therefore, the quantitative interaction between the identified risk factors specified in our model tends to be more realistic and accurate.

5. CONCLUSIONS

In this paper we have presented a machine learning approach to a secondary analysis of data collected in a college drinking study. The proposed machine learning approach can effectively identify risk and/or protective factors for high-risk drinking that might be effective in detecting early developmental signs of alcohol abuse or dependence within college-aged students. We demonstrated the use of a statistical feature selection method, 1-norm (sparse) support vector machine, to help classify college students as either heavy or low-risk drinkers and simultaneously select the risk factors for heavy drinkers. Additionally, our approach was shown to be accurate by the included ROC curve discussion and would perform well on future data sets. Our analyses demonstrate that support vector machine works well in classifying students into drinking behavior categories and is able to identify both protective and risk factors. The discussed results also offer new insights and further support existing hypotheses and theories onto motives for alcohol consumption of college students.

6. ACKNOWLEDGMENTS

We thank our collaborator and consultant, Dr. Victor Hesselbrock at UConn Alcohol Research Center for his insightful discussions.

7. REFERENCES

[1] US Department of Health and Human Services, "Healthy people 2010," Washington, DC, 2000.

[2] Office of Juvenile Justice and Delinquency Prevention, "Drinking in america: Myths, realities and prevention policy," in *Department of Justice, Office of Justice Programs*, Washington, DC, 2005.

[3] A. C. Carter, K. O. Brandon, and M. S. Goldman, "The college and non-college experience: A review of the factors that influence drinking behavior in young adulthood," in *Journal of Studies on Alcohol and Drugs*, 2010, pp. 742–750.

[4] Alcohol Abuse and Alcoholism Task Force, "A call to action: changing the culture of drinking at u.s. colleges," in *Survey from National Institutes of Health.*

[5] E. M. Jellinek, "The disease concept of alcoholism," New Brunswick, NJ: Hilhouse, 1960.

[6] M. Windle and R. A. Zucker, "Reducing underage and young adult drinking: how to address critical drinking problems during this developmental period," in *Journal of Alcohol Research and Health*, January 2010, pp. 29–44.

[7] S. Armeli, T. S. Conner, J. Cullum, and H. Tennen, "A longitudinal analysis of drinking motives moderating the negative affect-drinking association among college students," in *Psychology of Addictive Behavior*, 2010, pp. 38–47.

[8] R. M. O'Connor and C. R. Colder, "Predicting alcohol patterns in first-year college students through motivational systems and reasons for drinking," in *Psychology of Addictive Behaviors*, 2005, pp. 10–420.

[9] M. P. Martens and et al., "Understanding sport-related drinking motives in college athletes: psychometric analyses of the athlete drinking scale," in *Addictive Behaviors*, 2008, vol. 33, pp. 974–977.

[10] M. P. Martens and et al., "Do protective behavioral strategies mediate the relationship between drinking motives and alcohol use in college students?" in *Journal of Studies on Alcohol*, 2007, vol. 68, pp. 1–2.

[11] W. DeJong and et al., "NIAAA's rapid response to college drinking problems initiative: reinforcing the use of evidence-based approaches in college alcohol prevention," in *Journal of Studies on Alcohol and Drugs*, 2010, vol. 16, pp. 5–11.

[12] W. Dejong and et al., "A multisite randomized trial of social norms marketing campaigns to reduce college student drinking: A replication failure," in *Substance Abuse*, 2009, vol. 30, pp. 27–40.

[13] L. I. Zakletskaia and et al., "Alcohol-impaired driving behavior and sensation-seeking disposition in a college population receiving routine care at campus health services centers," in *Accident; Analysis and Prevention*, 2009, vol. 41, pp. 380–386.

[14] K. H. Beck and et al., "Social context of drinking and alcohol problems among college students," in *American Journal Of Health Behavior*, 2008, vol. 32.

[15] R. A. J. Singleton and A. R. Wolfson, "Alcohol consumption, sleep, and academic performance among college students," in *Journal of Studies on Alcohol and Drugs*, 2009, vol. 70, pp. 355–363.

[16] J. A. Orona and et al., "Examining drinking consequences and reasons for drinking in a bilingual college sample," in *Hispanic Journal of Behavioral Sciences*, 2007, vol. 39, pp. 101–115.

[17] J. W. LaBrie and et al., "Family history of alcohol abuse associated with problematic drinking among college students," in *Addictive Behaviors*, 2010, vol. 35, pp. 726–729.

[18] M. L. Cooper, "Motivations for alcohol use among adolescents: Development and validation of a four-factor model," in *Psychological Assessment*, 1994, pp. 117–128.

[19] C. C. Abar and J. L. Maggs, "Social influence and selection processes as predictors of normative perceptions and alcohol use across the transition to college," in *Journal of College Student Development*, 2010, pp. 496–508.

[20] M. P. Martens and et al., "Understanding sport-related drinking motives in college athletes: Psychometric analyses of the athlete drinking scale," in *Addictive Behaviors*, 2008, vol.33, pp. 974–977.

[21] J. D. Hawkins, R. F. Catalano, and J. Y. Miller, "Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for substance abuse prevention," in *Psychological Bulletin*, 1992, pp. 64–105.

[22] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," in *Machine Learning*, 2002, vol. 46, pp. 389–422.

[23] R. Yu and S. Shete, "Analysis of alcoholism data using support vector machines," in *BMC Genet*, 2005.

[24] B. Schölkopf, A. Smola, R. Williamson, and P. Bartlett, "New support vector algorithms," *Neural Computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[25] V. N. Vapnik, *The nature of statistical learning theory*. New York: Springer, 1995.

[26] A. P. Reynolds, G. Richards, and V. J. Rayward-Smith, "The application of k-medoids and pam to the clustering of rules," in *Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science*, 2004, pp. 173–178.

[27] J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song, "Dimensionality reduction via sparse support vector machines," *Journal of Machine Learning Research*, vol. 3, pp. 1229–1243, 2003.

[28] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (roc) plots: A fundamental evaluation tool in clinical medicine," in *Clin Chem*, 1993, pp. 561–577.

[29] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[30] R. A. Johnson and D. W. Wichern, "Applied multivariate statistical analysis 4th edition," Upper Saddle River, NJ, 1998.

[31] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," in *Journal of Machine Learning*, 2003, pp. 1157–1182.

[32] J. Wetson, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for svms," in *NIPS'13*, 2000.

[33] J. Wang, "Data mining: Opportunities and challenges," 2003.

[34] R. Kohavi and G. John, "Wrappers for feature subset selection (late draft)," in *Artificial Intelligence Journal, Special Issue on Relevance*, vol. 97, pp. 273–324.