A Quantitative Machine Learning Approach to Master Students Admission for Professional Institutions

Yijun Zhao
Computer and Information Science Department
Fordham University

Bryan Lackaye
Computer Science Department
Northeastern University

Jennifier G. Dy
Electrical and Computer Engineering Department
Northeastern University

Carla E. Brodley
Khoury College of Computer Sciences
Northeastern University

ABSTRACT

Accurately predicting which students are best suited for graduate programs is beneficial to both students and colleges. In this paper, we propose a quantitative machine learning approach to predict an applicant's potential performance in the graduate program. Our work is based on a real world dataset consisting of MS in CS students in the College of Computer and Information Science program at Northeastern University. We address two challenges associated with our task: subjectivity in the data due to change of admission committee membership from year to year and the shortage of training data. Our experimental results demonstrate an effective predictive model that could serve as a Focus of Attention (FOA) tool for an admission committee.

Keywords

support vector machine (SVM), semi-supervised learning, learning using privileged information (LUPI), multi-task learning, Educational Data Mining (EDM), business intelligence in education

1. INTRODUCTION

Master's education is the fastest growing and largest component of the graduate enterprise in the United States. According to the 2016 joint survey conducted by the CGS (Council of Graduate Schools) and ETS (Educational Testing Service) [4], first-time enrollment in U.S. graduate programs reached a record high total of 506,927 students in Fall 2015. Because of the rise in applicants, the admissions process may become increasingly tedious and challenging. The ETS has established standardized tests (such as the GRE) to help evaluate applicants' quantitative, reading, and writing skills, but these scores alone are far from indicative of successful students. Although applicants' previous achievements can demonstrate excellence, students with high GPAs from prestigious universities do not always excel in their graduate studies.

In this paper, we take a quantitative machine learning approach to predict the outlook of applicants' graduate studies based on features extracted from their application materials. The training data for our model are empirically admitted students with their performance measures in the graduate program. In particular, we have a real world dataset from Northeastern University's MS in Computer Science (MSCS) program, consisting of MS students from 2009 to 2012. We use a student's overall GPA in the MSCS program as his/her performance measure. Our model aims to identify the top 20% and bottom 20% performing students respectively (see details in Section 4.1).

Two challenges arise when learning with this data. First, the data involves the admission committee's (possibly subjective) evaluation. Specifically, some members of the committee may be biased in weighing a particular set of standards (e.g., GRE scores), while others may be in favor of different measures. This issue is particularly acute when the admission committee/policy changes from year to year. As a result, it can be difficult to form an accurate predictor directly from the entire dataset. Another challenge is the limitation of the training data. We have a total of 454 labeled training samples (all admitted students) from 2009 to 2012. On the other hand, we have over 2000 applications that are either rejected (i.e., not admitted) or declined (i.e., admitted but not enrolled), which can serve as an unlabeled auxiliary dataset. Our conjecture is that building a semi-supervised model leveraging the large set of unlabeled data may lead to a superior performance compared to using the labeled data alone.

Our model is inspired by two existing frameworks: SVM+ [12] and S3VM [3]. SVM+ is a variant of SVM which addresses the issue of heterogeneous data. Specifically, SVM+ implicitly establishes a different hyper-plane for each data subgroup by modifying a standard SVM's objective function and constraints. S3VM is a semi-supervised version of

Yijun Zhao, Bryan Lackaye, Jennifer Dy and Carla Brodley "A Quantitative Machine Learning Approach to Master Students Admission for Professional Institutions" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 538 - 544

SVM which learns a classifier using both labeled and unlabeled data. Our contribution is a new variant of SVM that unifies the advantages of both S3VM and SVM+. Our new model, which we name S3VM+, addresses the admission biases in the labeled data and utilizes unlabeled applicants' data simultaneously. S3VM+ can be applied to any domain for which the data may have clearly defined subgroups (e.g., privileged domain knowledge) and a large amount of unlabeled data.

An additional motivation of our research was to validate our hypothesis of whether we could predict student success based only on quantitative measures and, thus, remove the subjectivity of the committee reading the recommendation letters and statement of purpose. If successful, such a model will not only lead to a better selected student body, but also help to manage growing enrollments. Our experimental results (see Section 4 for details) demonstrate that, with our new model, we can achieve an effective yet imperfect prediction. Thus, in practice, our model could serve as a Focus of Attention (FOA) tool for the admission committees.

The rest of the paper is organized as follows: in Section 2, we present the related work in predicting students' performance in the education domain. In Section 3, we give brief introductions to S3VM and SVM+ and present our model S3VM+ in detail. We demonstrate the efficacy of our model in Section 4 by comparing its performance to those three existing models. Finally, we conclude in Section 5.

2. RELATED WORK

Most EDM studies focus on predicting students' academic performance after they have been admitted to the college or program. For example, Lepp et al. investigated the relationship between cell phone use and academic performance in a sample of US college students [8]. Delen applied machine learning techniques for student retention management [6]. Ioanna et al. presented a dropout prediction in e-learning courses using machine learning techniques [10]. Nevertheless, another important aspect of educational research is selecting the best fitting students at admission time, which has not been widely addressed in past literature.

The most closely related work to our paper is the admissions research conducted by the University of Texas at Austin (UT Austin) for their graduate admission program [14], driven in part by their need to manage growing application numbers. In their work, the authors applied logistic regression (LR) to help the admission committee identify weak candidates who will likely be rejected and exceptionally strong candidates who will likely be admitted. Our work bears a similar mission but is different in three aspects. First, the UT Austin research includes credentials such as recommendation letters and statement of purpose, whereas our work strives to build a purely quantitative model relying only on non-subjective measures. Second, the recommendations made by UT Austin's algorithm are based on an applicant's likelihood of admission, whereas our model aims to predict the future performance of the applicants in the graduate program. Last, our model addresses human subjectivity in admission decisions. The contribution of our paper is a quantitative machine learning model to predict a candidate's future performance at admission time.

3. INTEGRATING SEMI-SUPERVISED SVM WITH DOMAIN KNOWLEDGE

We choose our model based on the characteristics of our dataset and particular challenges involved in our task. In particular, we choose SVM and two existing frameworks: S3VM [3] and SVM+ [12]), as our baseline models. Our proposed model is a new variant of SVM, which is inspired by S3VM and SVM+. We first give brief introductions to S3VM and SVM+. We then describe our new model in detail in Section 3.3.

3.1 S3VM (Semi-Supervised SVM)

S3VM is semi-supervised SVM proposed by [3]. The model is learned using a mixture of labeled data (the training set) and unlabeled data (the auxiliary set). The objective is to assign class labels to the auxiliary set such that the "best" support vector machine (SVM) is constructed. In particular, given a labeled dataset $L = \{x_1, x_2, \ldots, x_l\}$ and an unlabeled auxiliary dataset $U = \{x_{l+1}, x_{l+2}, \ldots, x_{l+k}\}$, S3VM learns a classifier from both L and U using overall risk minimization (ORM) posed by Vapnik [13] (Chapter 10). Starting with the standard SVM formulation, S3VM adds two constraints for each data point in the auxiliary set U. One constraint calculates the misclassification error as if the point were placed in class 1, and the other constraint calculates the misclassification error as if the point were placed in class -1. The objective function calculates the minimum of the two possible misclassification errors. The final membership assignments of the instances in U correspond to the ones that result in a minimum total sum of slacks across all instances in the training set. Specifically, we have:

$$\min_{w,b,\eta,\xi,z} \quad \frac{1}{2} \parallel w \parallel^2 + C \left[\sum_{i=1}^{l} \eta_i + \sum_{j=l+1}^{l+k} min(\xi_j, z_j) \right] \quad (1)$$

subject to

$$\begin{array}{lll} y_i(w \cdot x_i + b) + \eta_i \geq 1 & \eta_i \geq 0 & i = 1, \ldots, l \\ w \cdot x_j + b + \xi_j \geq 1 & \xi_j \geq 0 & j = l+1, \ldots, l+k \\ -(w \cdot x_j + b) + z_j \geq 1 & z_j \geq 0 & j = l+1, \ldots, l+k \end{array}$$

where C is the trade-off between maximizing the margin and total violations. η_i's are the slacks for the labeled data, and ξ_j's and z_j's are the slacks for the unlabeled data hypothetically assigned to the positive and negative classes respectively.

Equation (1) can be solved using mixed integer programming by applying the "large integer M" technique. The idea is to introduce a constant integer $M>0$ and a decision variable $d_j \in \{0, 1\}$ for each point x_j in the auxiliary set U. d_j indicates the class membership of x_j. If $d_j=1$, then the point is in class 1 and if $d_j=0$, then the point is in class -1. The integer M is chosen sufficiently large such that if $d_j=0$ then $\xi_j=0$ is feasible for any optimal w and b. Likewise if $d_j=1$, then $z_j=0$. In other words, ξ_j and z_j can have at most one non-zero value no matter what class x_i belongs to. Consequently, we could replace the $min(\xi_j, z_j)$ in Equation (1) by

$(\xi_j + z_j)$. This results in the following formulation:

$$\min_{w,b,\eta,\xi,z} \quad \frac{1}{2} \parallel w \parallel^2 + C \left[\sum_{i=1}^{l} \eta_i + \sum_{j=l+1}^{l+k} (\xi_j + z_j) \right] \tag{2}$$

subject to

$$\begin{aligned} & y_i(w \cdot x_i + b) + \eta_i \geq 1 \\ & \eta_i \geq 0, \quad i = 1, \ldots, l \\ & w \cdot x_j + b + \xi_j + M(1 - d_j) \geq 1 \\ & -(w \cdot x_j + b) + z_j + M d_j \geq 1 \\ & \xi_j \geq 0, \quad z_j \geq 0, \\ & j = l+1, \ldots, l+k, \quad d_j \in \{0, 1\} \end{aligned}$$

The solution to Equation (2) can be found using mixed integer programming products. In our experiment, we used CVX [1] and Gurobi [2] optimizers. Same as a standard SVM, S3VM classifies a new instance x^* using $sign(w * x + b)$.

3.2 SVM+

Vapnik and Vashist [12] introduced SVM+, which is a variant of SVM that addresses the issue of learning with heterogeneous data. In their model, the authors developed a new paradigm to learn using privileged information (LUPI). The objective of SVM+ is to take advantage of additional domain knowledge, and in particular data subgroups that may arise from different sources or due to labeling biases.

Suppose the training data has $t > 1$ groups. We follow the notation in [9] and denote the indices of group r by

$$T_r = \{i_{n_1}, \ldots, i_{n_r}\}, \quad r = 1, \ldots, t$$

All training samples can then be represented as:

$$\{\{X_r, Y_r\}, \quad r = 1, \ldots, t\}$$

where $\{X_r, Y_r\} = \{(x_{r_1}, y_{r_1}), \ldots, (x_{r_{n_r}}, y_{r_{n_r}})\}$. To incorporate the group information, SVM+ defines the slacks inside each group by a unique *correcting function*:

$$\xi_i = \xi_r(x_i) = \phi_r(x_i, w_r), \qquad i \in T_r, \quad r = 1, \ldots, t$$

Specifically, the correcting functions are defined as:

$$\xi_r(x_i) = w_r \cdot x_i + d_r, \qquad i \in T_r, \quad r = 1, \ldots, t$$

Compared to a standard SVM, S3VM uses slack variables that are restricted by the correcting functions, and the correcting functions capture additional information about the data. Note that all of the data is used to construct the decision hyperplane. The group information is only used to fine tune the slack variables. Formally, the objective function for $SVM+$ is formulated as follows:

$$\min_{\substack{w,b,w_1,w_2,w_r,\\ d_1,d_2,d_r}} \quad \frac{1}{2} \parallel w \parallel^2 + \frac{\gamma}{2} \sum_{r=1}^{t} \parallel w_r \parallel^2 + C \sum_{r=1}^{t} \sum_{i \in T_r} \xi_i^r \tag{3}$$

subject to:

$$\begin{aligned} & y_i(w \cdot x_i + b) + \xi_i^r \geq 1 \\ & \xi_i^r(x_i) = w_r \cdot x_i + d_r \\ & \xi_i^r \geq 0, \quad i \in T_r, \quad r = 1, \ldots, t \end{aligned}$$

Parameter γ adjusts the relative weight between $\parallel w \parallel^2$ and the $\parallel w_r \parallel^2$'s. C is the trade-off between maximizing the margin and total violations.

Liang and Cherkassky [9] further extended the SVM+ approach to multi-task learning. In the SVM+MTL [9] framework, the data is partitioned into groups using privileged information similar to the SVM+ model. However, instead of making a correcting function for the slack variables, their model establishes a unique correcting function (i.e., a hyperplane) for each group in addition to a shared common hyperplane. In other words, the decision function for group $r = 1, \ldots, t$ is as follows:

$$f_r(x) = (w \cdot x + b) + (w_r \cdot x + d_r)$$

where w, b are the parameters for the common hyper-plane and w_r, d_r are the parameters for the correcting function for group r. The corresponding formulation of the quadratic optimization problem is as follows:

$$\min_{\substack{w,b,w_1,w_2,w_r,\\ d_1,d_2,d_r}} \quad \frac{1}{2} \parallel w \parallel^2 + \frac{\gamma}{2} \sum_{r=1}^{t} \parallel w_r \parallel^2 + C \sum_{r=1}^{t} \sum_{i \in T_r} \xi_i^r \tag{4}$$

subject to

$$\begin{aligned} & y_i[(w \cdot x_i + b) + (w_r \cdot x_i + d_r)] + \xi_i^r \geq 1 \\ & \xi_i^r \geq 0 \quad i \in T, \quad r = 1, \ldots, t \end{aligned}$$

SVM+MTL is an adaptation of SVM+ for solving MTL problems. In our experiment, we applied the SVM+MTL framework because it provides more flexibility to learn a different decision plane for each year's student data.

For SVM+MTL, predicting the class label for a new given instance x^* is not straightforward because its decision function requires a group-dependent correcting function, and we do not know the group membership of test instances. To resolve this problem, we predict the label for x^* in each group and perform a majority vote over all predicted labels. Specifically, a test instance x^* will be predicted in each group as follows:

$$f_r(x^*) = sign[(w \cdot x^* + b) + (w_r \cdot x^* + d_r)]$$

where $r = 1, \ldots, t$ are the bias groups, and w, b, w_r's and d_r's are learned model parameters. The class membership for x^* is determined by a majority vote over $f_r(x^*)$'s.

3.3 S3VM+

Our new model, S3VM+ leverages the unlabeled data and addresses the biases in the training data simultaneously. In particular, we train our model with a labeled dataset and an unlabeled auxiliary dataset. Furthermore, our data is partitioned into yearly groups because of the admissions committee changes from year to year and thus may have different biases. For the labeled dataset, we incorporate the grouping information by establishing a correcting function for each group (constraints (a) and (b) in Equation (5)).

For the unlabeled data, we introduce two slack variables ξ_i and z_i for each data point x_i representing the slacks of placing x_i in the positive class and negative classes respectively. The objective function for S3VM+ takes the minimum of the two slacks for each unlabeled instance and minimizes the total sum of slacks across all training instances. We apply the "large integer M" technique (see Section 3.1 for details) and convert the constraint with a minimization function to two constraints over linear functions. Because both labeled and

unlabeled data are grouped by academic year, we apply the same correcting functions used for the labeled data to each corresponding annual group of unlabeled data (constraints (c) to (f) in Equation (5)). Formally, the optimization problem for S3VM+ is formulated as follows:

$$\min_{w,b,w_1,w_2,w_r,d_1,d_2,d_r} \quad \frac{1}{2} \parallel w \parallel^2 + \frac{\gamma}{2} \sum_{r=1}^{t} \parallel w_r \parallel^2 + C \left[\sum_{i=1}^{l} \eta_i^r + \sum_{j=l+1}^{l+k} (\xi_j^r + z_j^r) \right] \tag{5}$$

subject to

(a) $y_i[(w \cdot x_i + b) + (w_r \cdot x_i + d_r)] + \eta_i^r \geq 1$
(b) $\eta_i^r \geq 0 \quad i = 1, \ldots, l$
(c) $[w \cdot x_j + b + (w_r \cdot x_j + d_r)] + \xi_j^r + M(1 - d_j) \geq 1$
(d) $\xi_j^r \geq 0 \quad j = l+1, \ldots, l+k, \quad d_j \in \{0, 1\}$
(e) $-[(w \cdot x_j + b) + (w_r \cdot x_j + d_r)] + z_j^r + M d_j \geq 1$
(f) $z_j \geq 0 \quad j = l+1, \ldots, l+k \quad d_j \in \{0, 1\}$

where C is the trade-off between maximizing the margin and total violations, and γ is the trade-off parameter between $\parallel w \parallel^2$ and the $\parallel w_r \parallel^2$'s. Note that constraints $(a), (b)$ are for labeled instances and constraints $(c) - (f)$ are for unlabeled instances.

To classify a new instance x^*, we follow the same approach as SVM+, which is to take a majority vote on class labels predicted by each group.

4. EXPERIMENTAL RESULTS

In this section, we first describe the process of constructing our training and testing dataset. We then discuss the methods we used to conduct our experiments in Section 4.2. We present our analysis of our experiments in Section 4.3.

4.1 Constructing the Training and Test Data

We have a real world dataset consisting of students from the MSCS program at Northeastern University. Table 1 presents the features we collected from students' applications for our experiment. Feature 1 contains the students' undergraduate GPAs adjusted according to each individual university's grading scale. For example, a 3.5 out of 5 and a 7 out of 10 would result in the same value. Feature 10 contains self-reported values representing the maximum number of lines of programming written by the student prior to joining the MS program. Feature 12 contains the rankings of the undergraduate institutions where the students obtained their bachelor's degrees. We classified the rankings into 4 categories with 1 being the most prestigious. The classification was performed manually according to the Best Global Universities list published by US News and World Report. The rest of the features are standardized test scores. Both the GRE and TOEFL had two versions of tests during 2009 - 2012 which use different scoring scales. Both of these tests are converted to their new versions of scoring scales using conversion tables provided by the ETS [4].

As mentioned in Section 1, our task is to identify successful candidates at the point of admission. One measure of success is MS-GPA in the MS program (as distinct from the input feature 1 "Undergraduate GPA"). Indeed, a cumulative MS-GPA is the most widely used measure for students' academic performance [11]. The labels in our training data are determined by the training instances' percentiles in the overall MS-GPAs. Specifically, the top and bottom 20% students are labeled with class 1 and -1 respectively. The number 20% was intuitively chosen as an measure which sets the individuals apart from the average students.

Table 1: Features Collected for Training

1	Undergraduate GPA
2	GRE Verbal
3	GRE Quantitative
4	GRE Analytical Writing
5	TOEFL Total
6	TOEFL Reading
7	TOEFL Listening
8	TOEFL Speaking
9	TOEFL Writing
10	Max # of Lines of Code Written
11	Bachelor's Degree in EECS (Yes/No)
12	Undergraduate School Ranking

Table 2: Student Data Statistics

Year	Total	Top 20%	Bottom 20%	Aux. Data
2009	37	7	7	431
2010	89	18	17	503
2011	132	28	27	705
2012	196	51	42	948

Note that we did not use a midpoint MS-GPA as a cutoff to separate the positive and negative classes, in order to reduce the label noise. In particular, instances close to the average GPA are harder to categorize as good or bad students.[1] Another intuitive approach is to define two hard MS-GPA thresholds for good versus bad performances, i.e., to have a MS-GPA above an upper threshold (e.g., > 3.8) for good students, and below a bottom threshold (e.g., $<$ 3) for bad students . A further investigation reveals that this approach is less effective for the following reason: different instructors have different grading policies due to the nature of the courses. For some fundamental courses, an 'A' means you are in the top 30% of a class, while for some other advanced courses, an 'A' means you are in the top 10% of a class. Even for the same course in the same year with different sections, the instructors may choose to cooperate exams/grading or not. Because students have different instructors and/or even take different courses, hard cutoffs are not an accurate reflection of a student's abilities.

Having stated this, on the other hand, if a student performs consistently in the top 20% in each class, this student will be among the top 20th percentile of the entire MS-GPA spectrum. The same can be said for those that perform consistently in the bottom 20th percentile. Identifying the factors that lead to this consistent success or underperformance are of greatest interest to this research. Therefore, we used relative measures to label our positive and negative training samples. For comparison purposes, we report our experimental results on both relative and hard cutoffs in Tables 5 and 6 respectively.

[1]We did experiment with splitting the two classes using the mean value of all MS-GPAs and the performance was not satisfactory as expected.

Table 3: Prediction Using 1Y Data

Train	Teset	Top20% MS-GPA	Bot20% MS-GPA	Overall
2009	2010	0.72	0.59	0.66
2010	2011	0.64	0.70	0.67
2011	2012	0.65	0.76	0.70

Table 4: Predicting Using 10-fold Cross Validation

	Test Accuracy			Training Accuracy		
MS-GPA% / Model	Top20	Bot20	Overall	Top20	Bot20	Overall
2009 - 2011	0.70	0.71	0.71	0.79	0.79	0.79
2009 - 2012	0.74	0.72	0.73	0.84	0.75	0.79

Table 2 summarizes the distribution of students from 2009 to 2012. Column "Total" is the total number of students enrolled in the corresponding year. Columns "Top 20% MS-GPA" and "Bottom 20% MS-GPA" are the total number of students in the top and bottom 20th percentile among their peers measured by the cumulative MS-GPAs . There is not an equal number of positive and negative instances for each year because there are multiple students with same MS-GPA.

Both SVM+ and our model S3VM+ make use of an unlabeled auxiliary dataset. We collect the application data of rejected (i.e., not admitted) and declined (i.e., admitted but not enrolled) applicants as the auxiliary data. These data contain the same features as the labeled data, and the size distribution of auxiliary data from 2009 to 2012 is presented in the last column of Table 2. Our training data are all labeled and unlabeled instances from 2009 to 2011, and our test data are labeled instances from 2012.

4.2 Experimental Method

We are interested in identifying the top and bottom 20% of candidates from an application pool based on the performance of the admitted students. Our first goal is to confirm our conjecture that there are biases in admission decisions from year to year. To this end, we conducted two experiments. The first experiment is to use the previous year's data to predict the current year's performance using a standard SVM. For example, we would use class 2009's data to predict class 2010's performance, and class 2010's data to predict class 2011's performance. Table 3 presents the prediction accuracies for each year. We observe that, for 2010, the top 20% of students are easier to predict than the bottom 20%, whereas for 2011 to 2012, the situation is reversed. This lack of consistency and the low overall accuracies (up to 70%) suggest that there is no strong correlation of predicative patterns from year to year. Our second experiment is to apply a standard 10-fold cross validation on two datasets: all data from 2009 to 2011 and all data from 2009 to 2012. Because 2012 added a significant amount (89%) of instances, we would expect a noticeable increase in both the training and test accuracies if the data across different years conform to the same distribution. Table 4 summarizes the results of this experiment. We observe only a marginal improvement in overall test accuracy after adding instances from 2012 and, more importantly, the overall fit of the data remains the same (79%). From these two experiments, we conclude that data across different academic years have different distributions. We believe this year to year bias is due to the change in the membership of the admission committee.

Table 5: Performance Comparison with Relative Cutoffs

	Test Accuracy			Training Accuracy		
MS-GPA% / Model	Top20	Bot20	Overall	Top20	Bot20	Overall
SVM	0.73	0.71	0.72	0.79	0.80	0.79
S3VM	0.75	0.74	0.74	0.81	0.82	0.81
SVM+	0.77	0.70	0.74	0.92	0.84	0.88
S3VM+	**0.82**	**0.72**	**0.77**	**0.95**	**0.89**	**0.92**

Table 6: Performance Comparison with Hard Cutoffs

	Test Accuracy			Training Accuracy		
MS-GPA / Model	>3.8	<3.4	Overall	>3.8	<3.4	Overall
SVM	0.65	0.69	0.66	0.73	0.75	0.74
S3VM	0.72	0.65	0.70	0.83	0.70	0.77
SVM+	0.75	0.64	0.71	0.92	0.75	0.84
S3VM+	**0.77**	**0.67**	**0.74**	**0.93**	**0.80**	**0.87**

In light of above learned information, we partitioned the data by academic year and use them as the privileged groups in SVM+ and S3VM+. We take the union of labeled data from 2009 to 2011 as our labeled training data. The auxiliary dataset is formed as the union of the corresponding auxiliary data from 2009 to 2011. We test and compare the performance of the four models (SVM, S3VM, SVM+, S3VM+) in predicting labeled instances in 2012.

The hyper-parameters are the trade-off constant C for all four models and γ for SVM+ and S3VM+. We perform 10-fold cross validation and grid search on the training data to select the hyper-parameters. We first use a coarse grid $\{0.01, 10, 1000\}$ for C and refine the candidates after the initial search. The final list for C is $\{1, 10, 100\}$. Following a similar procedure, our final search list for γ is $\{0.01, 1, 100\}$. After the best hyper-parameters are selected, we train the corresponding model one more time using the entire training data and then apply the learned model to the test data and measure its performance. We report both training and test accuracies in Table 5.

4.3 Analysis on Performance Measures

Table 5 displays the main results of our experiment. First, we observe that the test accuracies for SVM on the positive and negative classes are more balanced compared to the results in Table 3. There is also an improvement in the overall performance for SVM. This can be explained by the increased amount of training data used in our Table 5 experiment.

Second, we conclude that all three variants of SVM (S3VM, SVM+, S3VM+) are superior to standard SVM. Using SVM as a baseline measure:

- S3VM improved slightly on the accuracies of both positive and negative classes, which suggests that using auxiliary data has a positive impact on identifying both the good and bad students. This is consistent with the fact that the auxiliary data contain both de-

clined (i.e., admitted but not enrolled) and rejected (i.e., not admitted) applicants, which could improve the accuracy of positive and negative classes respectively.

- SVM+ demonstrated improvement on the positive side only, which indicates that the partition of bias groups by academic year is most effective in identifying the top students. One explanation for this could be that the top 20% of students are inherently different from year to year, while the bottom 20% of students remain similar. Or that a particular admissions committee has biases about how to recognize a strong student.

- Our model S3VM+ has a noticeable advantage among all models in predicting the positive class: 83% versus 73% (SVM), 75% (S3VM) and 77% (SVM+). In light of the construction of S3VM+, one could conclude that adding auxiliary data to each partition group further enhances the power of identifying top students. On the other hand, because grouping does not have a significant impact on identifying bottom students (as demonstrated by SVM+), S3VM+ would only result in a limited gain for the negative class.

Lastly, from the training accuracies presented in Table 5, we observe a significantly better fit of the training data using our model S3VM+. In particular, 95% versus 92% (SVM+), 81% (S3VM), 79% (SVM) accuracies for the positive class and 89% versus 84% (SVM+), 82% (S3VM) and 80% (SVM) accuracies for the negative class. Compared to the standard SVM, S3VM improved training accuracies evenly on both classes, and SVM+ and S3VM+ demonstrated more significant gains on the positive class, which is consistent with what we observed in the test data.

4.4 Labeling Strategy: Relative v.s. Absolute

Recall that in Section 4.1, we discussed our choice of labeling the top 20% and bottom 20% of students with respect to their MS-GPAs as our two classes. We explained our rationale of using relative rather than hard cutoffs to label our data. We confirm this conjecture in Table 6, where we show the results of an experiment using MS-GPA > 3.8 for the top students and MS-GPA < 3.4 for the other. In the table we see that for all four methods, the overall accuracies are lower than in Table 5.

4.5 Analysis on Weight Vectors

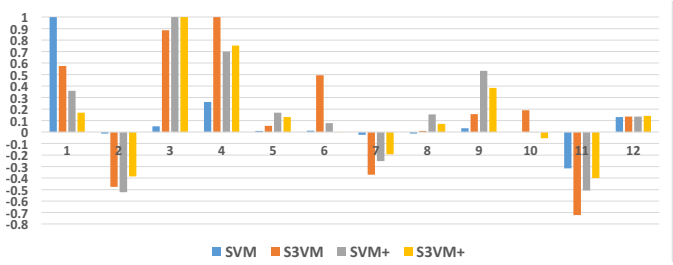
Because we utilized a linear SVM and its variants, we found it interesting to investigate the ranking and magnitude of each individual feature in the weight vectors produced by each model. Table 7 presents the ranking of w_i's in the weight vectors (w's) of four models. Figure 1 displays the weights of individual features across four models using their magnitudes. In order to make a meaningful comparison, each weight vector $w = \{w_1, w_2, \ldots, w_{12}\}$ is scaled by the maximum absolute value of its components. Thus, the weight for the most important feature is either 1 or -1. Note that, for SVM+ and S3VM+, we display the shared hyper-plane vector w without the correcting functions for each group.

From Table 7, we observe that all models except standard SVM suggest the same top two features: "GRE Quantitative" and "GRE Analytic Writing" scores. Furthermore,

Table 7: Weights Ranking Comparison

Model	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
SVM	w_1	w_{11}	w_4	w_{12}	w_3	w_9	w_7	w_2	w_6	w_8	w_5	w_{10}
S3VM	w_4	w_3	w_{11}	w_1	w_6	w_2	w_7	w_{10}	w_9	w_{12}	w_5	w_8
SVM+	w_3	w_4	w_9	w_2	w_{11}	w_1	w_7	w_5	w_8	w_{12}	w_6	w_{10}
S3VM+	w_3	w_4	w_{11}	w_2	w_9	w_7	w_1	w_{12}	w_5	w_8	w_{10}	w_6

Figure 1: Weights Distribution Over 12 Features Across Four Models



The weights are normalized using $w = \frac{w}{\max_{1 \le i \le 12} \{|w_i|\}}$

SVM+ and S3VM+ overlapped in their top five features but with a different ranking order.

From Figure 1, we conclude that the most important features are 1 ("Undergraduate GPA"), 2 ("GRE Verbal"), 3 ("GRE Quantitative"), 4 ("GRE Analytical Writing") and 11 ("Bachelor's Degree in EECS (Yes/No)"). A closer examination reveals that SVM relies mostly on three features (1, 4, and 11). S3VM has significantly large weights on two additional features, 6 ("TOEFL Reading") and 7 ("TOEFL Listening"), on top of the five features listed above. SVM+ and S3VM+ made use of one additional feature which is 9 ("TOEFL Writing").

5. CONCLUSIONS

In this paper, we applied a quantitative machine learning approach to predict candidates' potential academic performances based on information from their applications. We built our model using empirically admitted students with their cumulative GPAs as performance measures and tested our model's efficacy for the incoming students. Throughout our experiments, we found a unique challenge associated with our task, which is different data distributions across the academic years due to biases arising from changing membership of the admissions committee. We addressed this issue with the Learning Using Privileged Information (LUPI) framework. We further handled the limited training data issue by employing a semi-supervised version of SVM to utilize the large amount of unlabeled data (i.e., the rejected/declined applications). Our resulting model, S3VM+, is a novel variant of SVM that addresses subjectivity and lack of labeled data simultaneously. Our experimental results demonstrate a significant gain of our model compared to three existing models in standard literature (i.e., standard SVM, S3VM, and SVM+). Although we based our work on a two-year master's program, our model is easily extensible to similar tasks such as college or pre-school admissions. Our model can also be applied to other real world situations in which data may have clearly defined biased subgroups and a large amount of unlabeled data.

6. REFERENCES

[1] *CVX Research Inc.* www.cvxr.com.

[2] *Gurobi Optimizer*. www.gurobi.com.

[3] K. Bennett and A. Demiriz. Semi-supervised support vector machines. *NIPS*, 11:368–374, 1998.

[4] CGS and ETS. Graduate enrollmentand degrees: 2005 to 2015. *http://cgsnet.org/reports*, 2016.

[5] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20.3:273–297, 1995.

[6] D. Delen. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49.4:498–506, 2010.

[7] EDM. International educational data mining society. *http://www.educationaldatamining.org/*, 2016.

[8] A. Lepp, J. E. Barkley, and A. C. Karpinski. The relationship between cell phone use and academic performance in a sample of us college students. *Sage Open*, 5.1:2158244015573169, 2015.

[9] L. Liang and V. Cherkassky. Connection between svm+ and multi-task learning. *IEEE World Congress on Computational Intelligence*, pages 2048–2054, 2008.

[10] I. G. V. N. G. M. Lykourentzou, Ioanna and V. Loumos. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers and Education*, 53, no. 3:950–965, 2009.

[11] A. M. Shahiri and W. Husain. A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72:414–422, 2015.

[12] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22.5:544–557, 2009.

[13] V. N. Vapnik. *Estimation of dependences based on empirical data*. New York: Springer-Verlag, 1982.

[14] A. Waters and R. Miikkulainen. Grade: Machine learning support for graduate admissions. *AI Magazine*, 35.1:64, 2014.