Rose-Hulman Undergraduate Mathematics Journal

Volume 1
Issue 1

Article 1

Optimization Methods Applied to and Compared Through an Academic Database

Lori Giles
University of Evansville

Optimization Methods Applied to and Compared Through an Academic Database

Lori Giles
lgiles@math.utk.edu
University of Evansville
Undergraduate Research Project
Summer 1998

Abstract

We present results and conclusions stemming from the application of various optimization methods on an academic database. The goal is to provide a tool for our client to use to predict the best prospective students based on data gathered pre-registration. We applied several optimization programs to our database. The methods will be compared and contrasted based on accuracy and transferability of results to future student data. We also analyze the general "goodness" of the database itself, and propose possible improvements that will aid in better classification.

1. Introduction

Academic institutions are businesses. An important goal is to successfully educate students and prepare them for the job market in their chosen fields. Academic institutions are interested in students who are successful and will finish their degrees at that institution. It would be helpful to colleges and universities if they could efficiently predict during the admission process which students will successfully graduate. Consequently, this determination will be completed by using only high school performance and personal information given by each student. This information generally includes GPA, SAT scores, ACT scores, class rank, household income, age, address, activities the student was involved in, and work experiences. These attributes will vary from one institution to another. Our goal is to provide a predictor, or cut off criterion, for the University of Evansville to help them choose prospective students more efficiently.

We applied several existing mathematical programming techniques for classification and compared these results to the traditional statistical method, logistic regression. These methods include the mathematical programming method RLP, introduced by Bennett and Mangasarian [2]. RLP minimizes the average magnitude of the misclassification errors. Other mathematical programming methods we applied stem from RLP. These methods are RLP-P [1], and FM RLP-P [5] by Bennett and Bredensteiner. These methods incorporate into their model improved generalization by minimizing the maximum classification error and/or minimizing the number of attributes or features from the data set used. The traditional statistical method of logistic regression [14] implemented on SPSS will be used as a comparison. A brief

discussion of these methods is included in Section 2. In Section 3, the specifics of the classification problem given to us will be discussed and analysed. Section 4 contains the results we obtained using the above mentioned methods and how they can be interpreted to help the University of Evansville.

1.1 Notation

The following notation is used. Let $\mathcal{A}^1$ and $\mathcal{A}^2$ be two sets of points in the n-dimensional real space $\mathcal{R}^n$ with cardinality m_1 and m_2 respectively. Let A^1 be an $m_1 \times n$ matrix whose rows are the points in $\mathcal{A}^1$. Let A^2 be an $m_2 \times n$ matrix whose rows are the points of $\mathcal{A}^2$. Let e denote a vector of ones of the appropriate dimension. The scalar 0 and a vector of zeros are both represented by 0. Thus, for $x \in \mathcal{R}^n$, $x > 0$ implies that $x_i > 0$ for $i = 1,...,n$. For the column vector x in $\mathcal{R}^n$ and the matrix A in $\mathcal{R}^{n \times m}$, the transpose of x and A are denoted x^T and A^T respectively.

2. Classification Methods

The classification problem considered is the discrimination between the two sets $\mathcal{A}^1$ and $\mathcal{A}^2$ in n-dimensional real space, $\mathcal{R}^n$. Each dimension of the space represents an attribute of the elements of the sets. The function that is generated to separate these sets must not only correctly classify the points it was constructed on but also correctly classify future data. In this paper we use linear classification functions, such as a line in $\mathcal{R}^2$ or a plane in $\mathcal{R}^3$. There are also methods available that generate non-linear classification functions [6, 9, 16] but we have limited our investigation to linear functions.

Generally, classification involves determining a linear function that consists of a linear combination of the attributes of the given sets. The most basic problem is one in which a linear function, called a perceptron, can be used to completely separate the two sets. (Figure 1.0)

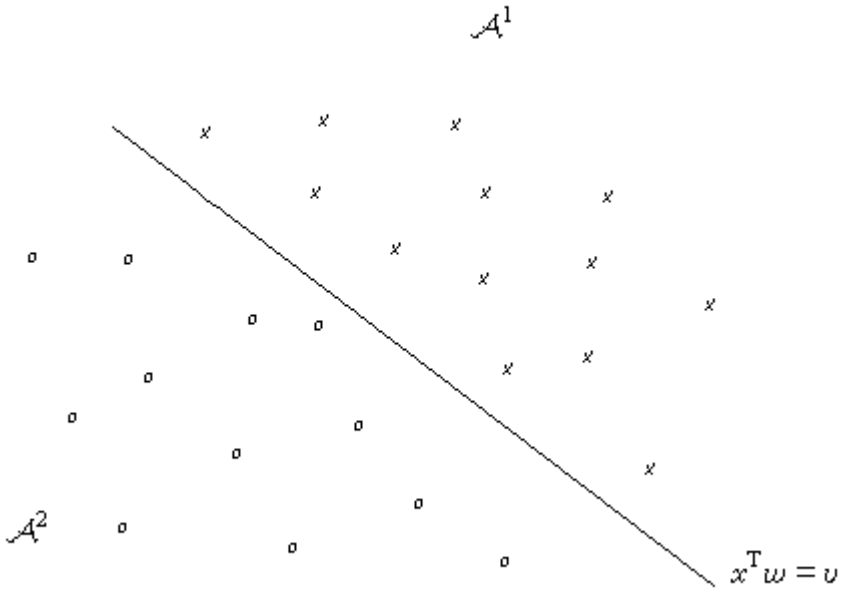

Figure 1.0 Example of separable sets and their perceptron

This function determines a separating plane, where all of the points from one set are on one side and all of the points in the other set are on the other side. Each x or o represents an n-dimensional vector, where each component, x_i, has a weight, w_i, associated with it. If the linear combination of these components with their weights, $\sum_{i-1}^{n} w_i x_i = x^T w$, is greater than some threshold, v, then the vector belongs to the set of x's in the graph. If however, the linear combination is less than the threshold, then the vector belongs to the set of o's, below the perceptron. Expanding this idea, let $\mathcal{A}^1$ and $\mathcal{A}^2$ be two sets in $\mathcal{R}^n$ with cardinality m_1 and m_2 respectively. Let A^1 be an $m_1 \times n$ matrix whose rows are the points of $\mathcal{A}^1$. Similarly, let A^2 be an $m_2 \times n$ matrix whose rows are the points of $\mathcal{A}^2$. Suppose x, an element of $\mathcal{R}^n$, is a point to be classified. A perceptron with weights w, in $\mathcal{R}^n$, and threshold v, in $\mathcal{R}$, would be defined as:

$$x^T w - v > 0 \quad \rightarrow \quad x \in \mathcal{A}^1$$

$$x^T w - v < 0 \quad \rightarrow \quad x \in \mathcal{A}^2$$

The preceding perceptron defines the separating plane where w is the normal to the plane and v determines the distance of the plane from the origin.

In general, the two sets of points, $\mathcal{A}^1$ and $\mathcal{A}^2$, are linearly separable if there exists w, v such that $A^1 w > ve$ and $ve > A^2 w$ where e is a vector of ones of the appropriate dimension. Upon normalization, this becomes $A^1 w - ve - e \geq 0$ and $-A^2 w + ve - e \geq 0$. As we might suspect, in

general it is not possible for a linear function to completely separate two given sets of points. Therefore, it becomes important to find the linear function that does the best job of separating the two sets. The *best* linear function is found by minimizing some error criterion involved in the incomplete separation. The following sections will discuss existing mathematical programming methods used to find this best function.

2.1 Robust Linear Program (RLP) and Perturbed Robust Linear Program (RLP-P)

RLP [2] and RLP-P [1] are mathematical programming methods that find a linear function that separates two sets of data. They are unique among other programs in the process they use to find this best function. RLP and RLP-P minimize the average magnitude of misclassification errors (See Figure 1.1).

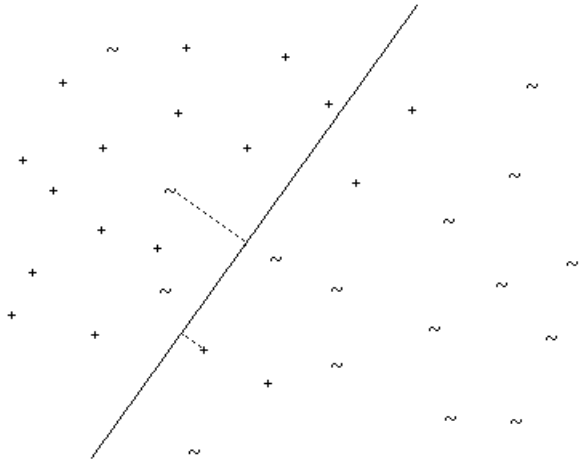

Figure 1.1 Two linearly inseparable sets. The average of the distances of the misclassified points is minimized when using RLP and RLP-P.

The subsequent robust linear programming (RLP [2]) problem is the problem solved to find the best linear function in this manner.

$$\begin{aligned} \min_{w,v,y,z} \quad & \tfrac{1}{m_1} e^T y \; + \; \tfrac{1}{m_2} e^T z \\ subject\ to \quad & z - A^2 w + ve - e \geq 0 \\ & y + A^1 w - ve - e \geq 0 \\ & y \geq 0,\ z \geq 0 \end{aligned}$$

The vectors y and z contain the elements y_i and z_j whose values are proportional to the distances of the misclassified points in sets $\mathcal{A}^1$ and $\mathcal{A}^2$ to the planes $A^1 w = ve + e$ and

$A^2w = ve - e$ respectively. Therefore, the objective function is the average of the distances of the misclassified points. Note that if the two sets are linearly separable then all y_i and z_j equal 0, and an infinite number of solutions is possible.

A modified version of RLP, called RLP-P [1], is introduced to find a classifier that will improve generalization on future points. In RLP-P, the average magnitude of misclassification errors is minimized and the distance between the supporting planes, $x^Tw = ve + e$ and $x^Tw = ve - e$, is maximized. (See Figure 1.2) The supporting planes are such that all of the points in $\mathcal{A}^1$ satisfy $x^Tw \geq ve + e$ and all of the points in $\mathcal{A}^2$ satisfy $x^Tw \leq ve - e$. By maximizing this distance, the resulting classification function $x^Tw = v$ is chosen to be as far away from both $\mathcal{A}^1$and $\mathcal{A}^2$ as possible. This distance is

$$\frac{2}{\|w\|}.$$

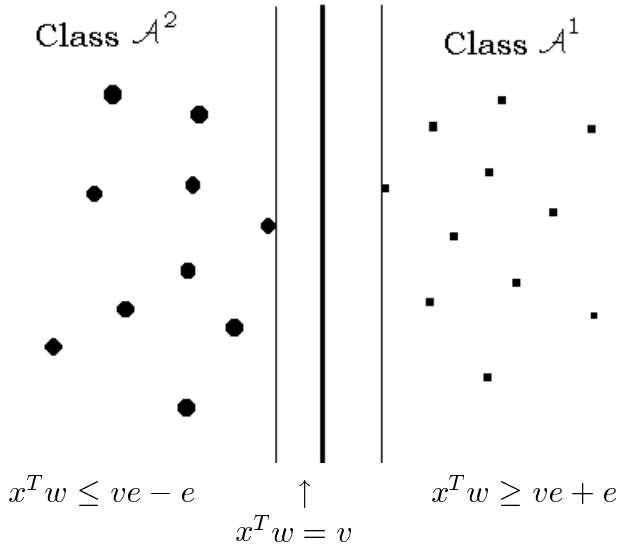

Figure 1.2 Separable sets and their separating planes

In order to maximize this distance, $\| w \| = e^T|w|$ is minimized. The absolute value would make the problem non-linear so it is removed by introducing a new variable, s. By adding the condition $-s \leq w \leq s$, the system is rewritten as a linear program. This new linear program is called RLP-P [1] or Perturbed Robust Linear Program and has the following form:

$$\begin{aligned} \min_{w,y,z\,s,v} \quad & (1-\lambda)\left(\tfrac{1}{m_1}e^Ty + \tfrac{1}{m_2}e^Tz\right) + \lambda e^Ts \\ subject\ to \quad & z - A^2w + ve - e \geq 0 \\ & y + A^1w - ve - e \geq 0 \\ & -s \leq w \leq s \\ & y \geq 0,\ z \geq 0,\ v \geq 0,\ s \geq 0 \end{aligned}$$

One must also note the added parameter λ in the objective function. λ is chosen by the user before running this program. It weights the two main pieces of the objective function. If for instance we want to focus more on minimizing the average magnitude of the misclassification errors, then we chose λ to be small. Conversely, if we want to focus on minimizing the 1-norm of w then we chose λ to be larger. Note that $0 < \lambda < 1$. For our purposes we chose λ to be 0.1 or 0.05.

2.2 Feature Minimization (FM RLP-P)

Feature Minimization [5] is another program we used on the university's database. Features or attributes are a type of information known for each piece of data. For example, for the university's database one feature was high school GPA. The goal of feature minimization is to construct good decisions or classification functions using as few features as possible while maintaining a certain level of accuracy. This is an important aspect of decision construction; especially when there is a large number of features involved. Minimizing the number of attributes cuts down on complexity, therefore making the results more understandable. Feature Minimization is useful when the user is interested not only in finding a successful classifier but also in using the output to better understand the fundamental characteristics of the data sets being studied. Feature selection methods have been employed successfully on a variety of data sets [4,8,11,12,13]

The unit step function, x_*, is used to count the number of nonzero elements in the vector w. The vector w is replaced with $(w_+) - (w\ )$ where $w_+,\ w\ \geq 0$. When the objective function is at optimality, the coordinates of the vectors w_+ and w satisfy $(w_+)_i = \max\{w_i, 0\}$ and $(w\ )_i = \max\{-w_i, 0\}$ for $i = 1, 2, ..., n$. Hence, the number of nonzero elements in the vector w is now $e^T(w_+ + w\ )_*$. The goal is to minimize this number while maintaining accuracy. When applying this technique to RLP-P, $e^T(w_+ + w\ )_*$ is added to the objective function and yields the following multi-objective optimization problem.

$$
\begin{array}{ll}
\min\limits_{w_+,\, w\ ,\, v,\, y,\, z} & [(1-\lambda)(\frac{1}{m_1}e^T y + \frac{1}{m_2}e^T z) + \lambda e^T(w_+ + w\ )] + e^T(w_+ + w\ )_* \\
\\
subject\ to & y + A^1(w_+ - w\ ) - ve - e \geq 0 \\
& z - A^2(w_+ - w\ ) + ve - e \geq 0 \\
& y \geq 0,\ \ z \geq 0,\ \ w\ \geq 0,\ \ w_+ \geq 0
\end{array}
$$

To improve ease of computing a solution, this problem can be reconfigured into a bilinear program. Following is the reconfigured, FM RLP-P, bilinear program we used on the university's database.

$$\min_{w_+, w^-, v, y, z, r} \quad (w_+ + w^-)^T(e - r)$$

$$\begin{aligned} subject\ to \quad & [(1-\lambda)(\tfrac{1}{m_1}e^T y + \tfrac{1}{m_2}e^T z) + \lambda e^T(w_+ + w^-)] \le \delta \\ & y + A^1(w_+ - w^-) - ve - e \ge 0 \\ & z - A^2(w_+ - w^-) + ve - e \ge 0 \\ & 0 \le r \le e, \quad e^T r \le \nu, \quad \nu \in [1, n] \\ & y \ge 0, \quad z \ge 0, \quad w^- \ge 0, \quad w_+ \ge 0 \end{aligned}$$

For more details on the intermediate steps consult *Optimization Methods in Data Mining and Machine Learning* by Erin Bredensteiner [7]. The value of δ used is $1.1[(1-\lambda)(\frac{1}{m_1}e^T y + \frac{1}{m_2}e^T z) + \lambda e^T(w_+ + w^-)]$, a ten percent increase of the optimal RLP-P objective function.

3. The Problem

The University of Evansville provided our data set. They are interested in studying retention using new techniques. Our goal was to find a classifier that would help them predict which students would actually graduate from UE given information they collected before the students' freshman year. Currently, the average percentage of students who graduate from the state of Indiana's private institutions is 46% [10].

We were provided with information from students who enrolled between the years of 1989 and 1993. The University had collected only ten pieces of viable information per student. This limited quantity of information may have hindered our ability to provide a successful classifier for them. The pieces of data they had collected include the following: social security number, gender, percent ranking in high school class, race, distance from home, SAT verbal score, SAT math score, SAT combined score, academic program code, and additionally indicated whether they graduated from the university. Whether they graduated or not is the class variable. If the SAT score was not given the university then calculated the SAT equivalent to ACT. The distance from home attribute is not measured in miles from UE. It was configured such that groups of states were given discrete values based on their distance from Indiana. These numbers also correspond with the percentage of students that generally come from those states to the University of Evansville. For instance, Indiana is assigned 1 and Tennessee, Illinois, and Ohio are each assigned 2. This trend continues as states move away from Indiana. It is necessary for all features to have numeric values. Therefore the race feature was configured similarly to the distance feature: a 1 was assigned to all Caucasian students, a 2 to all African American students, a 3 to all Asian students, and a 4 to the students who did not fall into any of the above groups. This ordering represents the most populous race to the least populous race represented at the university. These two attributes were defined in this way to reduce any problems with vagueness in ordering and to insure that the values had a logical and mathematical meaning. The program code attribute was restructured similarly to the distance attribute. Numerical codes were assigned to each program based on the student population of the program. The most heavily populated program was assigned the lowest

value and the least populated the highest. For this case, the range was chosen to be from five to fifty by fives.

4. Computational Results

The data set provided was classified using several programs including RLP-P, FM RLP-P, and SPSS's forward logistic regression procedure. The results of this problem will be reported in the form of gainscharts [3,15]. Gainscharts are used to glean the discriminating power of the classifier, by measuring the percentage of marked events taken at each decile. This information can then be used to make decisions about the given problem. Gainscharts show what kind of improvement we should expect if we choose or exclude groups of points, or in this case, students. The gainschart program we used lists the deciles such that the unsuccessful points are listed first, hence providing us data on the improvement expected if we deny admittance to the first decile of students, or the first and second deciles, etc. This way of looking at the results is most reasonable for our situation. The University of Evansville is not in a position to make large cuts in the number of admitted students. Therefore it is most useful to provide them a tool to show them the least desirable decile of the population. In this way they can be more selective without drastically reducing the campus population.

Our results also give some insight into which attributes are most crucial in the question of student success. As you will see in Table 2, FM RLP-P performed the best of the three methods presented. Interestingly enough, the decisions for this method were based purely on only two of our eight attributes, distance from home and rank percentile. All other attributes were not used, i.e. their coefficients were zero in the separating plane.

4.1 Gainscharts

The mathematical programming methods discussed were first run on a training set. During training, the coefficients for the best separating plane based on the student data are found. We then tested this separating plane using the testing set. These training and testing sets are simply obtained from the original data set. First we randomized the data and then broke it up into two sets. For the training set, we took 70% of the original data set. We want our training set to be large so the plane will be better developed. This is clear since it will be built using more information.

Once we obtained our classification functions, for the math programming methods we then rated the students in the following way. This is an approach similar to that in [3]. The distance each point, or student, is from the separating plane is calculated. The points that are in the target set are ranked in order of decreasing distance because the closer the point is to the separating plane, the closer it is to being misclassified. The points that are in the other set are ranked in order of increasing distance because the closer the point is to the separating plane the closer it is to being successful. Also, the closer the point is to the plane there is a greater chance it was misclassified. For the logistic regression results, we used SPSS to construct a linear classification function, $\beta_0 + \beta^T x$, where β, $x \in \mathcal{R}^n$ and $\beta_0 \in \mathcal{R}$. Forward stepwise selection was used to determine the set of significant features, i.e. the values β_i, $i = 1,2, \ldots, n$ are determined and the group of significant features correspond to the nonzero

components of β, whereas the insignificant features all have $\beta_i = 0$. The probability that a student is unsuccessful is calculated as

$$\frac{exp(\beta_0 + \beta^T x)}{1 + exp(\beta_0 + \beta^T x)}.$$

The students are then ranked in order of greatest probability of failure to least probability of failure.

After the points are ranked they are organized and the results are displayed in gainscharts. Each line of the gainschart contains 10% of the total population from the training set. The deciles appear in order of response rate. After this, the gainschart program keeps the same breaking points and divides the testing results into the existing deciles. This means that in the gainscharts presented, one decile does not necessarily represent 10% of the population. In our gainscharts the deciles are listed in order starting with the least successful students. It can also be arranged starting with the successful population but for our purposes, the data is more easily understood if it is listed least successful to most successful. In this way we can learn about the benefit of the consequence from cutting deciles of students. Each decile displays data about the total and target population; the target population being, in this case, the unsuccessful students. Percent of the total population, cumulative percent of the total population, cumulative percent of class 1 (the target set), cumulative response rate, and lift are included in the gainscharts. The gainscharts are given in Tables 1 through 3.

% Total Population in Node	Cumulative % Total Population	Cumulative % of Class 1	Cumulative Response Rate	Lift
8.72	8.72	10.30	57.69	118.02
7.94	16.67	20.37	59.73	122.20
10.29	26.96	32.95	59.75	122.24
8.39	35.35	41.42	57.28	117.18
8.61	43.96	50.11	55.73	114.00
10.51	54.47	61.10	54.83	112.16
9.62	64.09	69.57	53.05	108.54
9.51	73.60	77.57	51.52	105.40
11.97	85.57	89.24	50.98	104.29
14.43	100.00	100.00	48.88	100.00

Table 1: Results from using RLP-P

% Total Population in Node	Cumulative % Total Population	Cumulative % of Class 1	Cumulative Response Rate	Lift
24.72	24.72	30.43	60.18	123.12
22.04	46.76	52.86	55.26	113.06
10.51	57.27	63.16	53.91	110.28
9.96	67.23	72.08	52.41	107.22
9.96	77.18	80.32	50.87	104.07
9.96	87.14	89.24	50.06	102.42
9.96	97.09	97.48	49.08	100.40
2.91	100.00	100.00	48.88	100.00
0.00	100.00	100.00	48.88	100.00
0.00	100.00	100.00	48.88	100.00

Table 2: Results from using FM RLP-P

% Total Population in Node	Cumulative % Total Population	Cumulative % of Class 1	Cumulative Response Rate	Lift
7.83	7.83	9.15	57.14	116.90
8.17	16.00	18.54	56.64	115.88
8.72	24.72	29.29	57.92	118.49
9.17	33.89	39.13	56.44	115.45
10.74	44.63	48.74	53.38	109.21
6.60	51.23	59.29	53.71	109.88
8.61	59.84	65.22	53.27	108.98
9.62	69.46	74.60	52.50	107.39
12.42	81.88	85.58	51.09	104.52
18.12	100.00	100.00	48.88	100.00

Table 3: Results from using Logistic Regression

The first column of the gainschart, % Total Population in Node, represents the percentage of the population in each decile. Cumulative % Total Population is as one might expect, the percentage of the total population accrued up to a given decile. Similarly, Cumulative % of Class 1 is the percentage of the target class accrued up to the specified decile. Cumulative Response Rate is the percentage of the accumulated total population up to a given decile that is considered to be of the target class. The last column of the gainschart represents lift, a measure of how much better we are doing over choosing students at random. Lift is defined mathematically as:

$$\frac{100 (response\ rate)\ Class\ 1\ in\ population}{Class\ 1\ in\ decile}.$$

The bold lines in the charts show the predicted breaking point for deciles that contain points for which the majority of the decile population is the target class population. Recall that these breaking points, as well as the decile breaks, are based on the training set. Therefore they are meant to attempt to predict how the testing set will behave.

For example reading Table 1, it suggests we deny admission to the first and second deciles in order to exclude the majority of unsuccessful students. This would cut 16.67% of the students considered and deny admission to 20.37% of the target class. In addition, we also can read from the Cumulative Response Rate column that 59.73% of the 16.67% of the population cut should be in the target class, i.e. a majority of the population cut is in the target class. Finally the lift column tells us that if we make the proposed decision we will have improved by 22.20% over choosing students at random.

In comparing the three tables we can see that the FM RLP-P method reported the greatest lift when refusing admission to the first decile of students. However, eliminating the first decile would also mean eliminating 24.72% of the testing population, as indicated by the Cumulative % Total Population column (See Table 2). A major cut such as this in admissions may not be viable. Note that the other methods gave a lower lift value but are representing a much smaller cut in the population (See Tables 1 and 3).

Using the equation for the separating plane, information can be obtained concerning what features affect a student's possibility of graduating (See Table 4).

Method	**Attributes that Affect Graduation**
FM RLP-P	-Rank Percentile from High School -Distance from Home
RLP-P	-Rank Percentile from High School -Distance from Home -Race -SAT Combined Score
Logistic Regression	-Rank Percentile from High School -Distance from Home -Race -SAT Combined Score

Table 4 Attributes used for decisions in each method

These programs essentially choose the most influential attributes. Hence these results suggest that gender and academic program, for example, have no affect on a student's graduation potential.

Notice that the percentages in the gainscharts indicate intuitively that the methods we used are not necessarily the best for this problem. One might expect that these techniques would cause more than a 23% improvement over chance. The conclusion is that the data is not most successfully classified using a linear function. The average error that the programs reported was 35% for the training sets and 56% for the testing sets. Hence, such results are expected. Future work might be based on trying nonlinear classifiers such as polynomial classifiers, radial basis functions, or neural networks.

Additionally, to produce more useful results, the data set should be improved upon. By including additional attributes, such as amount of financial aid received, household income, number of activities involved in, the number of institutions each student applied to, and whether or not the University of Evansville was their first choice, the expanse of the information would be greater. These suggested features also can be logically connected to student success at an institution.

5. Conclusion

In conclusion, the methods we applied to the data set obtained from the University of Evansville produced informative results. These methods helped us predict what portion of the student body will be unsuccessful based on data from the students' high school careers and plans for study at the university. In addition we discovered what attributes are the most important with regards to student success. It appears that rank percentile from high school and distance from home are consistently valuable attributes but many more attributes are needed in order to gain better results.

Acknowledgements

Thanks to Manfred Schauss, Director of Institutional Research at the University of Evansville, for providing the data set used in this study.

Bibliography

[1] K.P. Bennett and E.J. Bredensteiner. Geometry in Learning. In C. Gorini, E. Hart, W. Meyer, and T. Phillips, editors, *Geometry at Work*, Washington, D.C., 1997. Mathematical Association of America. To appear.

[2] K.P. Bennett and O.L. Mangasarian. Neural network training via linear programming. In P.M. Pardalos, editor, *Advances in Optimization and Parallel Computing*, pages 56-67, Amsterdam, 1992. North Holland.

[3] K.P. Bennett, D.H. Wu, and L. Auslender. RPI Math Report No. 98-100, Rensselaer Polytechnic Institute, Troy, NY, 1998.

[4] P.S. Bradley, O. L. Mangasarian, and W. N. Street. Feature selection via mathematical programming. *INFORMS Journal on Computing*, 10:209-217, 1998

[5] E. J. Bredensteiner and K.P. Bennett. Feature minimization within decision trees. *Computational Optimization and Applications*, 10:111-126, 1998

[6] E. J. Bredensteiner and K. P. Bennett. Multicategory Classification by Support Vector Machines. *Computational Optimization and Applications*, 12:53-79, 1999.

[7] Erin Bredensteiner. Optimization Methods in Data Mining and Machine Learning. PhD Thesis. Rensselaer Polytechnic University, 1997.

[8] C. E. Brodley and P. E. Utgoff. Multivariate decision trees. *Machine Learning*, 19(1):45-77, 1995.

[9] C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:273-297, 1995.

[10] Independent Colleges of Indiana. 101 West Ohio St., Suite 440, Indianapolis, IN, 46204-1970.

[11] R. Kohavi and D. Sommerfield. Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 1995.

[12] O. L. Mangasarian. Machine learning via polyhedral concave minimization. In S. Schaeffler, H. Fischer, B. Riedmueller, editor, *Applied Mathematics and Parallel Computing - Festschrift for Klaus Ritter*, pages 175-188, Germany, 1996. Physica-Verlag. Technical Report 95-20, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, November 1995.

[13] O. L. Mangasarian. Mathematical programming in data mining. *Data Mining and Knowledge Discovery*, 1(2) 183-201, 1997

[14] I. Miller and M. Miller. *J.E. Freund's Mathematical Statistics*, 6th ed. Upper Saddle River, N.J.: Prentice Hall, 1999

[15] Ward Thomas. Database Marketing: Dual approach outdoes response modeling. Database Marketing News, page 26, June 1996.

[16] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.