Enrollment Prediction through Data Mining

Svetlana S. Aksenova, Du Zhang, and Meiliu Lu
Department of Computer Science
California State University
Sacramento, CA 95819-6021
lana.aksenova@intel.com, {zhangd, mei}@ecs.csus.edu

Abstract

In this paper, we describe our study on enrollment prediction using support vector machines and rule-based predictive models. The goal is to predict the total enrollment headcount that is composed of new (freshman and transfer), continued and returned students. The proposed approach builds predictive models for new, continued and returned students, respectively first, and then aggregates their predictive results from which the model for the total headcount is generated. The types of data utilized during the mining process include population, employment, tuition and fees, household income, high school graduates, and historical enrollment data. Support vector machines produce the initial predictive results, which are then used by a tool called Cubist to generate easy-to-understand rule-based predictive models. Finally we present some empirical results on enrollment prediction for computer science students at California State University, Sacramento.

Keywords: enrollment prediction, support vector machines, rule-based predictive models, Cubist.

1. Introduction

Enrollment prediction is of pivotal importance to the missions of a university [13]. Many methods have been proposed for enrollment prediction. They largely fall into the following two categories: (1) curve-fitting techniques such as simple or moving averages, exponential smoothing, polynomial or exponential models, and spectral analysis; and (2) causal models such as cohort-survival, ratio methods, Markov chain, multiple correlation and regression methods, path-analytical, and systems of equations [4, 14]. Recently, there have been efforts in utilizing some learning methods such as naïve Bayesian method and neural networks for enrollment prediction [15, 25, 29]. Different approaches build predictive models based on different types of data, target at different level of prediction (national, regional, state, university system, and institutional), have different error rates, and accomplish varying degrees of success.

Enrollment prediction depends on many factors [19]: social, economic and demographic trends, changes in technologies and job markets, university admission policies and program offerings, federal and state laws and regulations, costs and availability of financial aids and scholarships, locations and facilities/services, recruitment and advertising efforts, and many others. Though it is very difficult to sort out the influence of different variables whose movements are correlated over time, there is an agreement in the literature on which variables are reliably observed together to influence the enrollment trend. Through a broad range of studies, the variables below consistently prove to be statistically significant to enrollment [2, 27]: population change, family income, parent's level of education, tuition, student aid levels, and student's academic aptitude. It was found that the tuition fee level becomes less significant when family income increases [2]. There were also studies on the impact unemployment rates have on enrollment: higher unemployment rates result in higher college enrollment [2].

In this paper, we describe an enrollment prediction approach that is based on the following machine learning methods: support vector machines (SVM) [12, 23, 30] and rule-based predictive models. The goal is to predict the total enrollment headcount that is composed of new (freshman and transfer), continued and returned students at an institutional level. The proposed approach builds predictive models for new, continued and returned students at both undergraduate and graduate levels, respectively first, and then aggregates their predictive results from which the model for the total headcount is generated. The types of data utilized during the first step of the mining process include population and unemployment rates in the region, institutional tuition and fees, household income, high school graduates, and institutional historical enrollment data. Support vector machines produce the initial predictive results, which are then used by a tool called Cubist [21] to generate easy-to-understand rule-based predictive models.

0-7803-9788-6/06/$20.00 ©2006 IEEE.

The choice of predictive techniques in the study stems from the following consideration: the enrollment data are non-linear with complex boundaries and SVM offers an effective way in handling non-linear data, and rule-based predictive models are easy for human to comprehend.

The rest of the paper is organized as follows. Section 2 describes the types of data utilized in the study. Section 3 discusses the prediction methodology. Results of a case study are presented in Section 4. Finally, Section 5 concludes the paper with remark on future work.

2. Data sets

In addition to the historical enrollment data for a university, we choose to incorporate the following five types of data in the mining process: population, income per capita in constant dollars, tuition and fees, high school graduates, and unemployment rates. These five types of data represent a great social, economical and demographic influence on college enrollment and have been used in many similar studies.

Population. The population is continuing to grow and becomes more ethnically, and socioeconomically diverse. It seems reasonable to anticipate that the number of people enrolling from a given population group will be proportional to the size of that group. Work in economics and demography has emphasized that changes in the size of population groups may influence the rate at which members are likely to be involved in various activities, including enrolling in college. This is relevant to modeling the college enrollment decision [27].

Employment. There are reasons for including unemployment rates into enrollment models. There are offsetting effects of employment on enrollment, such as "discouraged worker" effect, where high school and university graduates who are unable to find work return to school [17]. An "added worker" effect where, when a parent is unemployed, children may not be able to afford to remain in school may impact enrollment. Similarly, students, who pay for their own education, may not be able to afford it as well. The work in [26] found that a cyclical unemployment variable is positively related to schooling in 8 of 9 cases and suggested the possibility that causality runs both ways between enrollment and unemployment. The work in [16] includes an unemployment variable in the female enrollment equations as an inverse measure of changes in the opportunity cost of obtaining a higher education. The variable turned out to be significant and positive, indicating that increased unemployment indicates a decreased opportunity cost, leading to higher female enrollment rates.

Tuition and Fees. Researchers have found a strong, inverse relationship between tuition fee levels and higher education enrollments [20]. Student fee increases, together with a shortfall in state financial aid grants, could have negative impact on student enrollment – especially for students from low-income and underrepresented groups. According to [7], a tuition fee increase of $100 results in an enrollment decline of between 0.5 and 1.0 percent. A study conducted for the Rand Corporation in 1995 showed that a 10 percent increase in student fees results in a 1.97 percent decline in California State University system enrollment [9]. A 2001 EdFund study showed that lower income students and students of ethnic groups other than white are more sensitive to tuition increases than students from middle and upper income backgrounds [8].

Income data. A higher household income suggests a greater ability to finance college education. Numerous research has been done to determine significance of income in enrollment prediction. The work in [17] finds that defining family income as after-tax household income, is positive and significant in three of five estimated models for males 18-19 and 20-24. The research in [1] included a variable measuring the median income of males aged 45-54 in their four age-sex models and found it positive and significant across the four groups. This same measure was used in another college enrollment model and was found to be positive and significant in all cases [27].

High School Graduates. High school graduates serve as an important predictor for estimating the first-time college freshman population. The California Department of Finance (DOF) Demographic Unit, the State's source of projected K-12 enrollments, annually provides projections of California's public high school graduates. DOF uses grade progression ratios derived from the most recent ten years of historical enrollment data from the Department of Education's California Basic Education Data System (CBEDS) database. DOF projects that the number of high school graduates will increase by 2010, but by 2012 it will start declining (Figure 1).

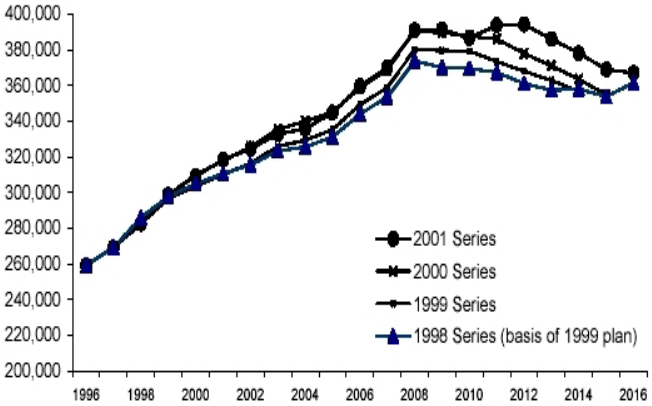

Figure 1. DOF Projections of California public high school graduates [28].

3. Prediction methodology

Our focus is to predict enrollments for a university that is situated in a certain social economical setting. Because an academic calendar starts at a fall semester (or a fall quarter), it makes sense to separate enrollment data into fall semesters and spring semesters (or fall, winter and spring quarters). Without loss of generality, we base our discussion on a two-semester scenario throughout the remainder of the paper.

For each semester, the total student headcount consists of the following:

1. New students: this includes the first time freshman students (high school graduates) and the transfer students from community colleges.
2. Continuing students: this group of students refers to those who attended the university in a previous semester (retention).
3. Returning students: those students who returned after some time away from the university or those who changed majors to enter the program (attrition).

Let RR, C, TE, and G denote "Retention Rate", "Continued", "Total Enrolled" and "Graduated", respectively, we have RR = C/(TE – G). RR may be different for fall-to-spring and spring-to-fall transitions. To predict the enrollment for the next fall from the current fall, we need to take this factor into consideration. (TE – G) is referred to as "Eligible to Continue" in the SVM model in Section 4.

After data are collected, our proposed approach consists of three steps. The first step concentrates on establish models for New, Continued, and Returned students, respectively. We use SVM on population and unemployment rates in the region, institutional tuition and fees, household income, high school graduates, and institutional historical enrollment data to produce the models.

The second step involves aggregating the results of New, Continued and Returned models to generate models for the total enrollment prediction through SVM.

In the third step, we utilize Cubist [21] to derive easy-to-understand rule-based predictive models to help gain insight into the enrollment prediction models. Figure 2 depict the mining process.

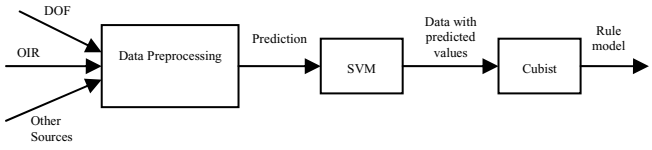

Figure 2. Generation of rule-based predictive models.

We use a support vector regression tool called DTREG [24] to generate SVM models. DTREG accepts a dataset consisting of number of rows and columns with a column for each variable. One of the variables must be identified as a "target variable" whose value is to be modeled and predicted as a function of "predictor variables". DTREG analyzes the data and generates a model, detecting how best to predict the values of the target variable based on values of the predictor variables. In addition to generating the predictive model, DTREG performs V-fold cross validation to measure the quality of the model.

Many kernel functions can be used for regression. In our approach, we choose the default and recommended kernel function of the Radial Basis Function (RBF) for a numeric prediction of New, Continued, and Returned enrollment. The settings for the parameter selection for the SVM models are kept the same for each initial dataset. To help gain understanding on the most important contributing factors, we enable the "Calculate variable importance" function of the tool so that it will generate a report on the relative significance of predictor variables for each model.

Cubist is a data mining tool that generates rule-based predictive models from data. Each rule is a multivariate linear model. When a situation matches a rule's conditions, the model is triggered to produce the predicted value.

We use the following absolute percentage error (APE) and mean absolute percentage error (MAPE) to evaluate the performance of the models.

$APE = |(predicted - actual)/actual|$

$MAPE = (\Sigma\ |(forecasted_i - actual_i)/actual_i|)/n$

4. Results of a case study

As a case study, we apply the approach to the enrollment prediction for computer science student headcounts at California State University, Sacramento (CSUS). In addition to the twenty-six years of population, income, high school graduates, and unemployment rates data for the state of California [5,6], the university tuition and fees, and historical computer science enrollment data are collected from CSUS' Office of Institutional Research (OIR) [11].

Since the historical enrollment data are organized by the OIR in two sequences of fall semesters and spring semesters at undergraduate and graduate levels, we first apply SVM to the data sets consisting of population, income, tuition fees, high school graduates and four-year college graduates, unemployment rates, and institutional historical enrollment to build component models for New (freshman and transfer), Continued and Returned computer science students, respectively. We use $New_u(f)$ and $New_g(s)$ to indicate "new undergraduate students in fall semesters" and "new graduate students in spring semesters", respectively. The same notation conventions apply to Continued and Returned students. Table 1

includes all the component models generated during the first step of the mining process.

Table 1. Component models.

	Undergraduate	Graduate
Fall semesters	- $New_u(f)$ - $Continued_u(f)$ - $Returned_u(f)$	- $New_g(f)$ - $Continued_g(f)$ - $Returned_g(f)$
Spring semesters	- $New_u(s)$ - $Continued_u(s)$ - $Returned_u(s)$	- $New_g(s)$ - $Continued_g(s)$ - $Returned_g(s)$

Each model generation starts with its initial data set. For instance, for $New_u(f)$, the initial data set includes six predictor variables: population, high school graduates, income, tuition, unemployment rate, and transfer (as university's historical enrollment data). Based on the variable (predictor) importance information generated by the SVM tool, different runs were made subsequently with the initial data set minus some less important predictor(s) to find a balance between the complexity of the model and the accuracy rate. For $New_u(f)$, the final model was generated based on four predictors (high school graduates, income, unemployment rate, and transfer) instead of the original six.

After all the component models were generated, we then use SVM to build the total headcount (tHC) models in terms of New, Continued and Returned predictions, for undergraduate and graduate levels, respectively. This is the second step in the mining process. Table 2 lists all the aggregated models. For instance, $tHC_U(f)$ was generated based on six predictors (Figure 3).

Table 2. tHC models.

	Undergraduate	Graduate
Fall semesters	$tHC_U(f)=\Sigma(New_u(f), Continued_u(f), Returned_u(f))$	$tHC_G(f)=\Sigma(New_g(f), Continued_g(f), Returned_g(f))$
Spring semesters	$tHC_U(s)=\Sigma(New_u(s), Continued_u(s), Returned_u(s))$	$tHC_G(s)=\Sigma(New_g(s), Continued_g(s), Returned_g(s))$

Finally, the total enrollment, THC, is defined as follows:

$THC(f) = tHC_U(f) + tHC_G(f)$,

$THC(s) = tHC_U(s) + tHC_G(s)$.

Due to space limit, we only include the SVM and Cubist models for $tHC_U(f)$ below. Interested readers may refer to [3] for a complete set of models.

```
Starting analysis at  5-Dec-2005 18:21:13
DTREG version 4.5  (Enterprise Demonstration
Version)
http://www.dtreg.com

  ============  Project Parameters  ============

Target variable: Total
Number of predictor variables: 6
Type of model: Support Vector Machine (SVM)
Type of SVM model: Epsilon-SVR
SVM kernel function: Radial Basis Function (RBF)
Type of analysis: Regression

  ============  Input Data  ============

Input data file: C:\Documents and
Settings\Owner\MyDocuments\MSPROJECT_12_4\TOTAL\
Undergrad\Fall\Ugrad_Fall1.csv
Number of variables (data columns): 9
Data subsetting: Use all data rows
Number of data rows: 40
Total weight for all rows: 40
Rows with missing target or weight values: 0
Rows with missing predictor values: 0

  =========  Summary of Variables  ============

Variable     Class      Type         Missing rows
-----------  ---------  -----------  ------------
Year         Unused     Categorical         0
Term         Unused     Categorical         0
Retention    Predictor  Continuous          0
Eligible_to
  _Continue  Predictor  Continuous          0
Graduated    Predictor  Continuous          0
Continued    Predictor  Continuous          0
New          Predictor  Continuous          0
Returned     Predictor  Continuous          0
Total        Target     Continuous          0


  ============  SVM Parameters  ============

Type of SVM model: Epsilon-SVR
SVM kernel function: Radial Basis Function (RBF)
SVM grid and pattern searches found optimal
 values for parameters:
  - Search criterion: Minimize total error
  - Number of points evaluated during search
     = 1191
  - Minimum error found by search = 0.112753

Parameter values:
  C = 392380.160233
  Gamma = 0.004644
  P = 0.064852

Number of support vectors used by the model = 30
  ==========  Analysis of Variance  ============

 ---  Training Data  ---

Mean target value for input data = 515.175
Mean target value for predicted values
   = 515.18038

Variance in input data = 27528.444
Residual (unexplained) variance after SVM model
   = 0.0177987
Proportion of variance explained by SVM model
   = 1.00000  (100.000%)

 ---  Validation Data  ---

Mean target value for input data = 515.175
Mean target value for predicted values
   = 515.23786

Variance in input data = 27528.444
Residual (unexplained) variance after SVM model
   = 0.1127528
Proportion of variance explained by SVM model
   = 1.00000  (100.000%)
    ……
  =====  Overall Importance of Variables  =====

      Variable          Importance
```

```
--------------------     ----------
Returned                    100.000
Continued                    51.609
New                          15.793
Eligible_to_Continue          0.013
Retention                     0.000
Graduated                     0.000

Finished the analysis at  5-Dec-2005 18:21:40
Analysis run time:  00:26.44
```

Figure 3. SVM model for $tHC_U(f)$.

In the last step of the mining process, Cubist models are generated from the data obtained through SVM models. The Cubist model for tHC_U for fall semesters is as follows.

```
Rule 1: [25 cases, mean 545.8, range 415 to 848,
         est err 8.5]
    headcount = 40.4 + 0.96 continued + 1.14 new
    Average error = 2.7%
    MAPE = 0.5%
```

Figure 4. Cubist model for $tHC_U(f)$.

Figures 5 and 6 indicate the predictions by the SVM and Cubist models for $tHC_U(f)$ and $tHC_G(f)$, respectively. Figure 7 is the scatter plot generated by the Cubist model for $tHC_U(f)$.

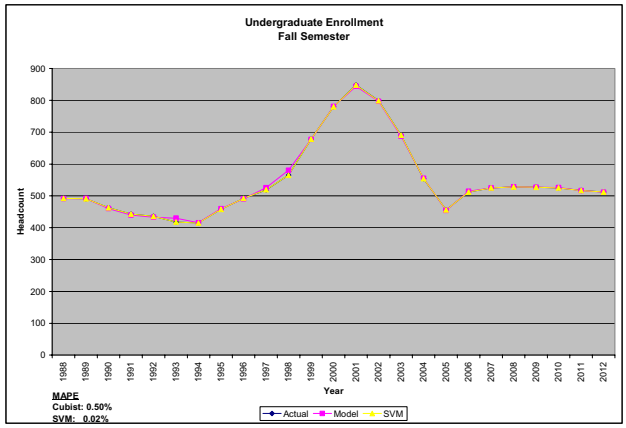

Figure 5. Predictions by SVM and Cubist models for $tHC_U(f)$.

The results demonstrate that both SVM and Cubist models did fairly well in terms of accuracy rate. The MAPE ranged from 0% to 11% for SVM models, and from 2.13% to 15.59% with average error up to 5.5% for Cubist models. These error rates are comparable to the results summarized in [2, 25].

Compared with the existing enrollment prediction methods, our approach has the following benefits:

SVM has greater generalization ability because of its structural risk minimization principle, and has performed well in many other forecasting applications [10, 18, 22]. SVM is adaptive to complex systems and robust in dealing with noisy data. We can easily add or remove variables in the mining process for the enrollment prediction. Furthermore, SVM does not suffer from the local minima problem that methods such as neural networks have.

The Cubist models afford those who do not have knowledge of data mining methods an opportunity to gain some insight on the significant factors in the prediction process.

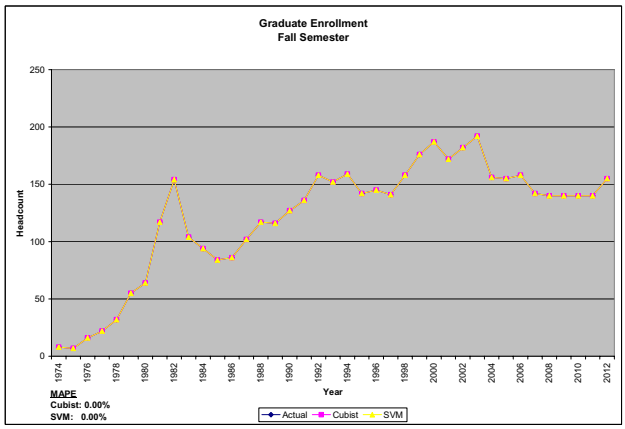

Figure 6. Predictions by SVM and Cubist models for $tHC_G(f)$.

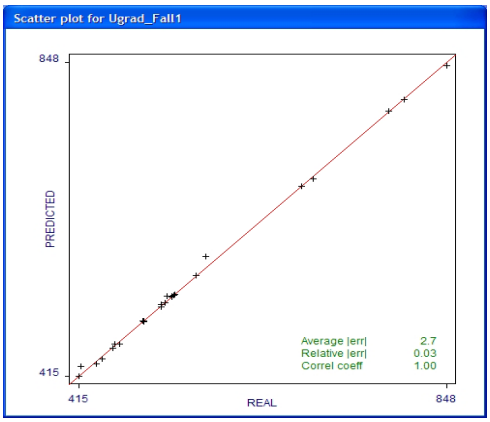

Figure 7. Scatter plot generated by Cubist model for $tHC_U(f)$.

5. Conclusion

Enrollment prediction is essential to a University's planning process. No matter how sophisticated the prediction models are or how large the data set is, there are limitations in enrollment projections. Selection of predictors sensitive to enrollment is not an easy task. There are emerging factors and conditions, and changing trends.

Data mining offers an effective type of predictive data analysis and has tremendous applications in higher education. As part of the future work, we plan on applying the approach to additional enrollment data from

other disciplines, examining and incorporating some additional predictors internal and external to a university into the mining process. We will also look into an ensemble approach to the prediction problem.

References

1. D. Ahlburg, E. Crimmins, and R. Easterlin, "The Outlook for Higher Education: A Cohort Size Model of Enrollment of the College Age Population, 1948-2000," *Review of Public Data Use*, Vol.9, 1981, pp. 211-227.
2. D. Ahlburg, M. McPherson and M. O. Schapiro, "Predicting Higher Education Enrollment in the United States: An Evaluation of Different Modeling Approaches," Williams Project on the Economics of Higher Education, DP-26, August 1994.
3. S. S. Aksenova, *Enrollment Projection through Data Mining*, MS degree project, Department of Computer Science, California State University, Sacramento, Fall 2005.
4. R. L. Armacost and A. L. Wilson, "Three Analytical Approaches for Predicting Enrollment at a Growing Metro Research University," ERIC Database: ED 474 040, 2002.
5. California Department of Finance, http://www.dof.ca.gov/
6. California Department of Labor CALMIS.
7. The California State University, Business and Finance News, http://www.calstate.edu/BF/Newsletters/letters97-98/1097Issue.shtml, October 31, 1997.
8. The California State University Briefing on Student Fee Increases and Student Enrollment, http://www.calstate.edu/acadres/docs/Briefing-student_Fees_StudEnrolls_final.pdf.
9. CFA Research Brief #2: Student Fees, The Rand Corporation,. Institute on Education and Training. Santa Monica, CA, January 1995.
10. B. J. Chen, M. W. Chang, and C. J. Lin, "Load forecasting using support vector machines: a study on EUNITE competition 2001," report for EUNITE Competition for Smart Adaptive System. Available: http://www.eunite.org
11. Computer Science Reports, Office of Institutional Research, California State University, Sacramento, http://www.oir.csus.edu/Reports/FactBook/DEPT/CSC.cfm
12. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.
13. D. S. P. Hopkins, and F. M. William, *Planning Models for Colleges and Universities*, Stanford University Press, Stanford, 1981.
14. G. Kraetsch, "Methodology and Limitations of Ohio Enrollment Projections", *The AIR Professional File of The Association for Institutional Research*, No.4, Winter 1979-80.
15. J. Luan, Data Mining Applications in Higher Education, an Executive Report, http://www.spss.com/events/e_id_1471/Data%20Mining%20in%20Higher%20Education.pdf.
16. D. J. Macunovich, The Missing Factor: Variations in the Income Effect of the Female Wage on Fertility in the U.S., Unpublished, January, 1993.
17. J. P. Mattila, "Determinants of Male School Enrollments: A Time Series Analysis", *Review of Economics and Statistics*, 64(2), May 1982, pp. 242-51.
18. K. R. Muller, A. Smola, G. Ratch, B. Scholkopf, Kohlmorgen J., V. N. Vapnik, "Using support vector support machines for time series prediction," Image Processing Services Research Lab, AT&T Labs.
19. National Center for Educational Statistics, Projections of Educational Statistics to 2014, US Department of Education, Institute of Educational Statistics, NCES 2005-074, thirty-third edition.
20. G. Park and R. Lempert, *The Class of 2014. Preserving Access to California Higher Education*, Rand Monograph Report, 1998.
21. RuleQuest Research, http://www.rulequest.com/
22. D. C. Sansom, T. Downs, and T. K. Saha, "Evaluation of support vector machine based forecasting tool in electricity price forecasting for Australian National Electricity Market Participants," in Proc. Australasian Universities Power Engineering Conference, 2002.
23. B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
24. P. H. Sherrod, Classification and Regression Trees And Support Vector Machines (SVM) For Predictive Modeling and Forecasting, "DTREG" Software, http://www.dtreg.com.
25. Q. Song and B. S. Chissom, "New models for forecasting enrollments: Fuzzy Time Series and Neural Network Approaches", ERIC Database: ED 358 169, 1993.
26. M. L. Wachter and W. L. Kim, "Time Series Changes in Youth Joblessness," in R. Freeman and D. Wise, eds. *The Youth Labor Market Problems: Its Nature, Causes, and Consequences*, University of Chicago Press, 1982, pp.155-185.
27. M. L. Wachter, and W. L. Wascher, "Leveling the Peaks and Troughs in the Demographic Cycle: An Application to School Enrollment Rates," *Review of Economics and Statistics*, 66(2), May, 1984, pp. 208-15.
28. University of California Educating the Next Generation of Californians in a Research University Context: University of California Graduate and Undergraduate Enrollment Planning Through 2010, Planning and Analysis Academic Affairs Office of the President University of California, February 1999.
29. "Using Analytic Services Data Mining Framework for Classification: predicting the enrollment of students at a university – a case study," a Hyperion White paper, http://dev.hyperion.com/resource_library/white_papers/Data_Mining_WP.pdf.
30. V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, New York, 1995.