Deriving Financial Aid Optimization Models from Admissions Data

Le Van Thanh and Peter Haddawy
Asian Institute of Technology, Computer Science and Information Management Program
le.van.thanh@ait.ac.th, haddawy@ait.ac.th

Abstract - **This paper presents a novel approach to deriving probabilistic models that predict enrollment given applicant background and the amount of financial aid offered. Our Bayesian network models can be used to optimize various enrollment objectives. We present a novel efficient optimization algorithm that uses the models to maximize expected tuition revenue under capacity constraints including student-faculty ratio and accommodation. We demonstrate and evaluate our approach using four years of graduate admissions data from the Asian Institute of Technology, consisting of 7,788 applicants from 84 different countries. This data set is particularly challenging since reliable family income data is not available for students from most of these countries. Evaluating the Bayesian network model with 10-fold cross validation yields an ROC Az value of 0.8451, with a predictive accuracy of 82.70% at a threshold of 0.5. Comparing the results of the tuition revenue optimization model to the institute's current financial aid allocation practice shows that if single-term tuition revenue is the sole optimization criterion, the institute can achieve its current enrollment numbers while realizing significant savings in its financial aid budget. The prediction and optimization software is currently being incorporated into the institute's online admissions processing system.**

Index Terms – Bayesian network, data mining, enrollment management, financial aid optimization.

INTRODUCTION

Financial aid allocation is one of the most important resource allocation decisions universities make. Financial aid is used to achieve a number of enrollment objectives, including diversifying the student population, attracting strong students, and maximizing tuition revenue. While financial aid generally positively affects applicant enrollment decisions, the effect on the probability of enrollment varies across applicants. Therefore an accurate model of how financial aid affects enrollment decisions is crucial for effective allocation.

Previous research has addressed creation of models to understand the patterns of enrollment of prospects [1,2,3]. One attempt to use a statistical method called a probit model to compute an expected enrollment yield and an average discount was described in [1]. Using that model, the authors can predict the probability of enrollment of individual and a group of inquiries. The authors report a false negative error rate and false positive error rate of 76% and 5% respectively. In this paper, we will show an approach to alleviate this problem of imbalanced prediction results. Other machine learning techniques including decision trees, neural networks and logistic regression have been also used to find significant patterns of admitted applicants, predict enrollment yield and increase student persistence [2,3]. Willett [3] reported that he could predict re-enrollment using C5.0 (5-fold boosted), neural network and logistic regression with accuracy of 90%, 70% and 69% respectively. The evaluation was carried out on 100 randomly selected students by using 7 years of admission data. None of the previous work has related the result of enrollment prediction to financial aid optimization.

The problem of optimizing tuition revenue is similar to the problem of optimizing customer revenue that occurs in other industries. For example, airline passengers with reservations may either show up for the flight or not show up. Researchers have developed models to forecast the rate of cancellations and no-shows. Hueglin *et. al.* [4] used classification trees and logistic regression models to estimate the probability of a passenger to be a no-show as well as the probability that a passenger will cancel the booking prior to departure. Lawrence *et. al.* [5] applied a number of techniques, including C4.5, IBM ProbE and APMR to predict no-shows and then used the result for revenue analysis. They argued that revenue improvement can be gained ranging from 0.41% to 1.21% depending on the input excess demand.

The next section describes the methodology used to build the models to estimate the enrollment probability and the models to optimize tuition revenue. It is followed by evaluation of the accuracy of prediction and presentation and analysis of the results of optimization.

METHODOLOGY

The expected tuition revenue from any given applicant who has been offered a particular amount of financial aid can be obtained by multiplying the probability of enrollment by the revenue obtained at that financial aid level. As the financial aid increases, the probability of enrollment increases but the tuition revenue decreases. So for each applicant there will be a financial aid offer that maximizes the expected revenue from that student. Our objective is to offer each student the amount of financial aid that maximizes tuition revenue, subject to capacity constraints. Developing such an optimization model requires first developing a predictive model that can determine for any given student the probability of enrollment for each level of financial aid offered.

We use Bayesian networks to compute the probability of enrollment. A Bayesian network [6,7] is a graphical representation of a probability distribution. It is a directed acyclic graph in which nodes represent random variables and links represent probabilistic influences between the variables. Probabilistic dependence and independence are expressed by the presence or lack of paths between nodes in the graph. A central feature of Bayesian networks is their ability to reduce the problem of determining a large number of probabilities in the joint distribution – the probability of every possible event – to relatively few. There are exponentially many such probabilities, yet Bayesian networks achieve compactness by factoring the joint distribution into local, conditional distributions of each variable given its parents.

I. Data preparation

The data used in this research was collected from four years of graduate admission data at the Asian Institute of Technology, consisting of 7788 applicants from 86 different countries. Among 7788 applicants, 1438 applicants were admitted and enrolled in the institute.

Based on the available attributes of the admissions dataset and the information needed for the model, nine attributes were chosen as predictors for enrollment: age, marital status, nationality, institute of the previous degree, GPA of the previous degree, school, degree program (master or doctor) to which the applicant is applying, financial aid.

The attributes nationality and institute of the previous degree have large number of values (84 and 1707, respectively) without any intrinsic meaning. We thus decided to transform them into more meaningful values.

Because of the international environment in which AIT operates, it is not possible to request copies of income tax returns, which would provide strong evidence of applicants' financial resources. In lieu of this, we used the World Bank classification of countries according to their Gross National Income (GNI) to group countries into GNI categories of LIC (Low Income), LMC (Low Middle Income), UMC (Upper Middle Income), NOC (High Income, non-OECD) and OEC (High Income, OECD). Although GNI represents the wealth of the country, it can be used in the model as an attribute showing the statistical financial capability of the population coming from the same country.

The most important factor concerning the university of the previous degree is the quality of the academic programs there. We gauged this by correlating the GPA of the previous degree with that obtained at AIT. If students consistently enter AIT with somewhat average grades from a particular university but graduate with high grades, we take this as evidence of the high quality of that institution. We rated institutions on a scale from 0 to 10.

III. Bayesian networks for enrollment prediction

Experimentation with a number of different Bayesian network structures showed that the simple Naïve Bayes model in Figure 1 produced the best results. All variables are discrete except FINANCIAL_AID which is Gaussian.

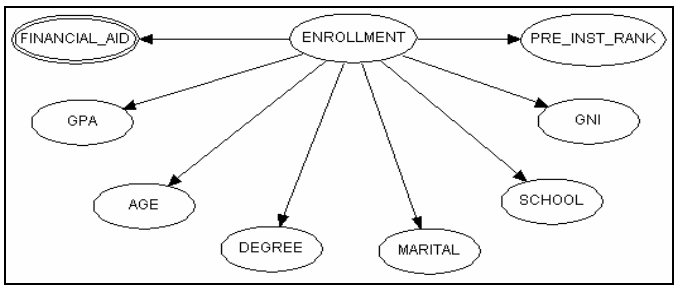


FIGURE 1
NAÏVE BAYES MODEL

The results of 10-fold cross validation are shown in Table I under Model 1. At a threshold of 0.5, the model has a predictive accuracy of 85.81%. But the performance is highly asymmetric, with a true negative rate of 95.62% but a true positive rate of only 42.49%. The reason for this is that the distribution of our admissions data is skewed. The number of examples of applicants who do not enroll greatly outnumbers the number of example of applicants who enroll: 82% and 18% respectively. When learning from such a skewed data set, machine learning algorithms tend to bias toward the majority class and do predict well for cases in majority class but badly for those in minority [9,10]. Another effect of the imbalanced dataset we found when performing financial aid optimization is that the model tends to recommend more generous financial aid offers than necessary because of the high false negative rate.

There has been much research on coping with the problem of imbalanced data sets [9] and various solutions have been proposed. At the data level, random over-sampling and random under-sampling [10] are popular techniques. At the algorithm level, many approaches have been proposed such as SMOTE [12,13] and DataBoost-IM [14].

Based on the experience on credit card fraud detection [11,15] and our large data set which consists of two possible uncertainties, we neither did under-sample majority examples which can remove important cases nor did over-sample minority cases which can cause over fitting problems. Instead, we split the majority class into a number of partitions such that we get the desired distribution when combining each split partition with examples in minority class. We experimented with learning the parameters of the Naïve Bayes model from the data sets synthesized from the original one with various minority:majority distributions and found the model trained on data with distribution of 60:40 biased towards positives cases produced the highest predictive accuracy for positive cases. This is shown as Model 2 in Table I.

TABLE I
ACCURACY INDICATORS FOR MODEL 1 AND MODEL 2

Model name	AUC	Accuracy	Enroll	True Positive/ Negative Rate	False Positive/ Negative Rate
Model 1	0.846	85.81	Yes	42.49	4.38
			No	95.62	57.51
Model 2	0.828	55.61	Yes	90.47	52.28
			No	47.72	9.53

We can gain more insight into the predictive models by examining the histograms of negative and positive predictions produced by each model to see how much the two distributions overlap. We would like models that produce well separated distributions. Figure 4 shows that the probability distribution of positive cases is spread over too much of the spectrum for Model 1 and Figure 5 shows a similar problem with negative cases for Model 2.

The complementary nature of the two models suggests the use of an ensemble model to combine the strengths of both. Ensemble techniques can generate a new machine learning scheme which is usually much more accurate than the individuals [8,17]. There are many algorithms developed to combine classifiers such as bagging [8], boosting [16] and stacking [8]. In this case, we use stacking to derive a new meta-classifier that can help us to decide in which cases we should trust model 1 and in which cases we should trust model 2. The meta-classifier of our stacking model is another Bayesian network with the structure shown in Figure 2. The meta-classifier consists of predicted ENROLLMENT variable and four continuous variables which are the outputs of model 1 and model 2.

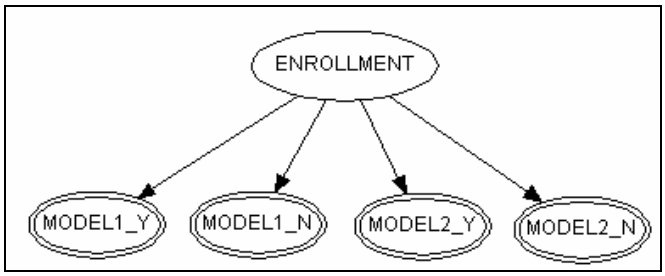


FIGURE 2
META-CLASSIFIER FOR STACKING

When the stacked learner is used for prediction, a new instance is first fed into the model 1 and model 2, each of which will produce a probability of enrollment. These predictions are then fed into the meta-classifier, which combines them into final prediction.

We evaluated the accuracy of this meta-classifier over the entire original imbalanced data set which consists of 7788 examples, yielding the ROC curve shown in Figure 3 and the performance indicators shown in Table II.

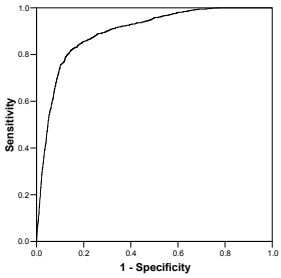


FIGURE 3
ROC CURVE OF THE META-CLASSIFIER

TABLE II
ACCURACY INDICATORS FOR META-CLASSIFIER

AUC	Accuracy	Enroll	True Positive/ Negative Rate	False Positive/ Negative Rate
0.8451	82.70%	Yes	65.79	13.46
		No	86.54	34.21

Compared with the model 1 and model 2, the stacked model has a much better balance between the true positive rate and true negative rate which are 65.79% and 86.54% respectively. In addition, the distributions of negative prediction and positive prediction (Figure 6) are now well separated.

IV Financial aid optimization model

The probability of enrollment computed by the meta-classifier can be viewed as a function of financial aid F and the background information B of the applicant: P(E|F,B). We can plot a curve reflecting the change in P(E|F,B) as we alter the percentage of financial aid. Figure 7 shows the curve relating probability of enrollment to financial aid for one particular applicant. While the probability of enrollment increases with an increase in financial aid offered, the relationship is typically nonlinear.

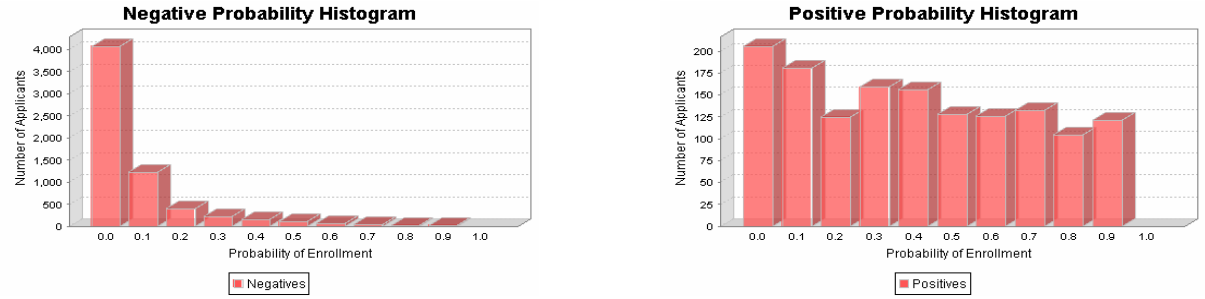


FIGURE 4
HISTOGRAMS FOR NEGATIVE AND POSITVE PREDICTION BY MODEL 1

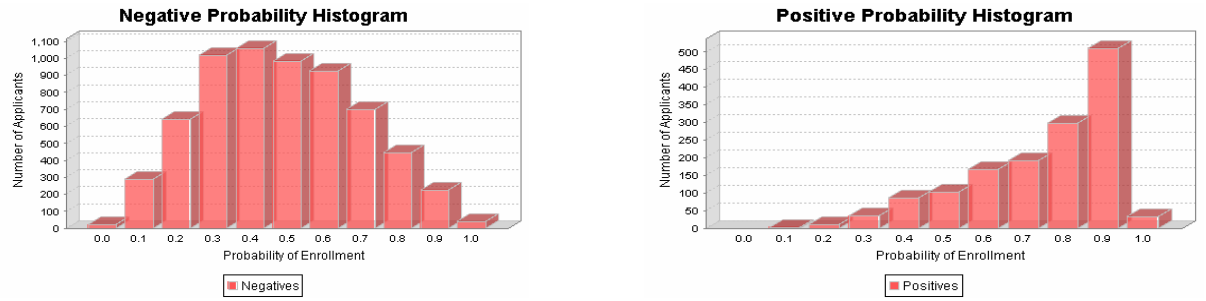


FIGURE 5
HISTOGRAMS FOR NEGATIVE AND POSITIVE PREDICTION BY MODEL 2

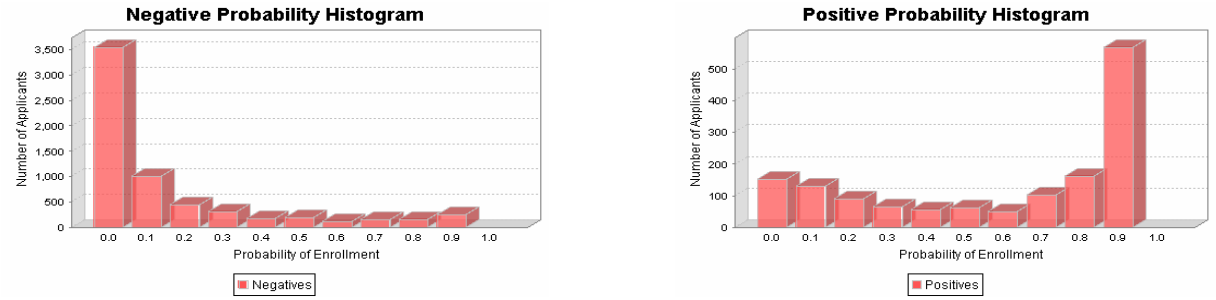


FIGURE 6
HISTOGRAMS FOR NEGATIVE AND POSITIVE PREDICTION BY META-CLASSIFIER

Figure 8 shows the simple linear relationship between percentage revenue and percentage financial aid, which is the same for all applicants. By multiplying the two curves, we can obtain a curve for the applicant that represents the expected revenue at each level of financial aid, as shown in Figure 9. Since the probability curve is monotonically increasing and the revenue curve is linear decreasing, the expected revenue curve is guaranteed to have a single maximum. For this particular applicant, the maximum expected revenue occurs when 50% financial aid is offered.

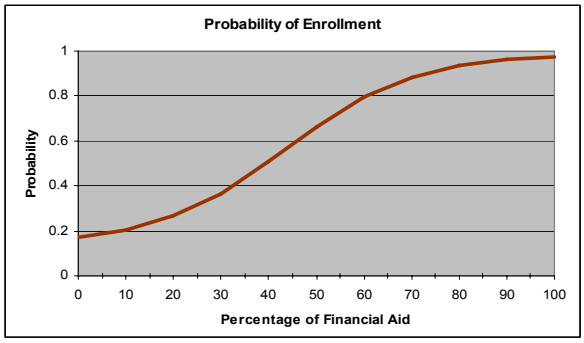


FIGURE 7
PROBABILITY OF ENROLLMENT GIVEN A PERCENTAGE OF FINANCIAL AID

In practice we cannot offer each applicant the amount of financial aid that would maximize revenue from that applicant individually because we must satisfy some capacity constraints, so this becomes a constrained optimization problem. We have global or institute-wide constraints such as housing capacity and local constrains such as the number of faculty members in each department combined with a maximum desirable student-faculty ratio. Hence the problem becomes determining the value F_i of financial aid for each student so that the institute receives the maximum total expected tuition revenue given these constraints:

$$\sum_{i=1}^{N} \arg\max_{F_i} P(E_i \mid F_i, B_i).T(F_i) \qquad (1)$$

where N is the total number of applicants offered admission

Since we have a single curve for each student with a unique maximum, this optimization problem can be solved with a simple greedy algorithm.

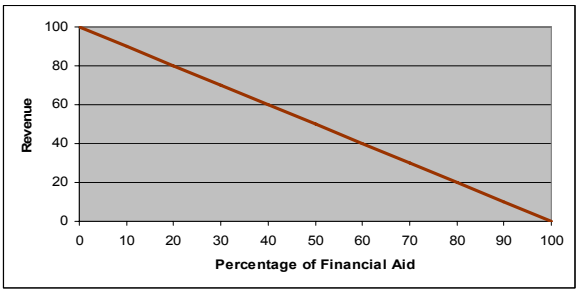


FIGURE 8
INSTITUTE'S REVENUE WHEN OFFERING A PERCENTAGE OF FINANCIAL AID

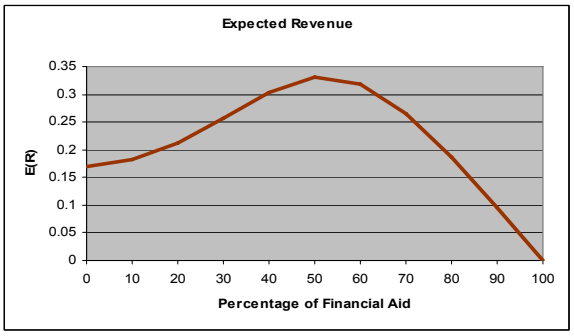


FIGURE 9
EXPECTED INSTITUTE'S REVENUE WHEN OFFERING A PERCENTAGE OF FINANCIAL AID

EVALUATION OF THE OPTIMIZATION MODEL

To evaluate the optimization algorithm, we used the 2006 admissions data from AIT consisting of 941 applicants who received between 0% and 100% financial aid from institute sources. Those who received scholarships from external donors were not included since the institute does not have control over these offers.

The algorithm was evaluated in two ways. First the algorithm was run without capacity constraints. This provides an indication of the maximum tuition revenue the institute could theoretically obtain from the existing pool of applicants. Then the algorithm was run with the constraint that the maximum number of students enrolling in each department equals the actual number that enrolled. Setting the constraint in this way highlights the cost savings that can result from using the optimization model.

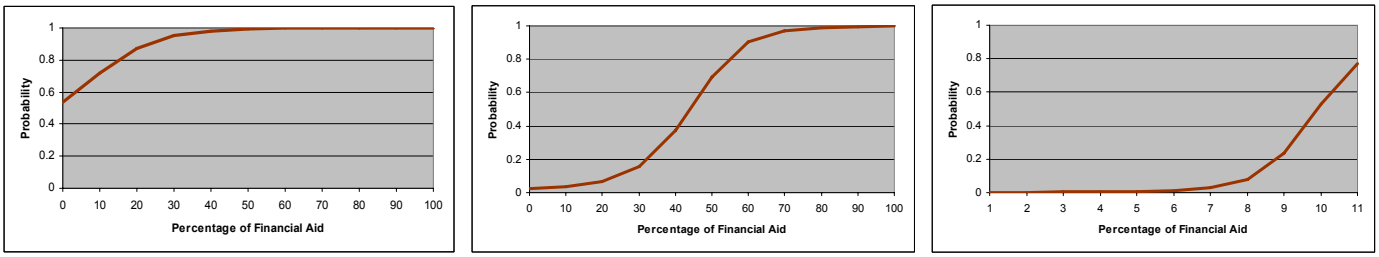
Running the algorithm without constraints, we obtained the profile of budget allocation as in the first row of Table III. Examining the results we find some interesting patterns among applicants. The probability curves for the applicants that fall into a given percentage of financial aid are very similar. Figure 10 shows typical curves for applicants in the percentage categories 20%, 60%, and 90%. Examining the attributes of the students in each of these groups reveals interesting characteristics. The applicants in the 90% group tend to come from poorer countries (GNI is LIC), have bachelor degree from a highly ranked institution (ranking in the range of 8 - 10), age over 24 and almost of them are married. The applicants in the 20% group tend to come from wealthier countries (GNI is LMC), have a previous degree from an institution with average ranking of 6 or 7, are single and quite young (less than 24). These results are in line with the experience of our admissions office.

Running the algorithm with the constraint that the maximum number of students recruited in each school is equal to the actual enrollment, we obtained the budget allocation as in the second row of the Table III. The result shows that those applicants receiving 80% or 90% financial aid in the unconstrained optimization now receive no financial aid since the institute has to spend too much to recruit these students. An interesting patterns lies in the applicants who the constrained optimization recommends providing 60% of financial aid. They tend to be single, graduate from universities with high rank, come from LIC countries, and are masters students. These results are consistent with the experience of our admissions officers.

Table IV shows an evaluation of enrollment prediction for each funding category. The table shows the number of applicants offered each amount of financial aid, the actual enrollment in that category, and the enrollment predicted by the model.

CONCLUSIONS

This paper has presented a novel approach to predict the probability of enrollment by using probabilistic networks. We also showed how to deal with the problem of imbalanced admission dataset, which is quite common in admissions data. Moreover, we demonstrated how to apply ensemble techniques to combine the strengths of multiple classifiers. Our optimization model currently considers only optimizing tuition revenue but financial aid is typically used to achieve other objectives, such as recruitment of strong students. Our model is general enough to admit such criteria and we are currently working on adding this capability to the implementation.



A – 20% FINANCIAL AID B – 60% FINANCIAL AID C – 90% FINANCIAL AID

FIGURE 10 (A-C)
RESPONSIVES OF GROUPS OF APPLICANTS TO THE OFFERED FINANCIAL AID

TABLE III
RESULTS OF EVALUATION OF FINANCIAL AID OPTIMIZATION ALGORITHM

Optimization Evaluation	Percent Financial Aid Offered										Total Number of Offers	Total Amount of Financial Aid Offered (USD)	Expected Enrollment	Expected Cost (USD)
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%				
Without constraints	52	11	41	22	46	120	231	135	233	50	941	11,666,000	442	5,316,000
With constraints	667	25	43	31	103	49	22	1	0	0	941	2,146,000	184	926,000
Actual	264	40	262	81	5	212	13	59	5	0	941	5,462,000	184	1,288,000

TABLE IV
ENROLLMENT PREDICTION EVALUATION

	Percent Financial Aid Offered										Total
	0%	10 %	20%	30%	40%	50%	60%	70%	80%	90%	
Total number of applicants	264	40	262	81	5	212	13	59	5	0	941
Expected enrollment	45.8	4.4	21.5	13.6	1.7	65	7.1	51.9	4.8	0	215.8
Actual enrollment	64	5	25	15	0	40	6	31	3	0	189

REFERENCES

[1] Martin R.E., Hokayem C.M., Leaf J. and Perry J., "Prospecting Among The Inquiries", *NACUBO Business Officer*, January 2002, 35-40.

[2] Chang L., "A Case Study: Applying Data Mining Technology in Modeling and Predicting Admissions/Enrollment Yield in Higher Education", *Association of Institutional Research 44th Forum Boston*, 2004.

[3] Willett T., "Opening the Black Box: How Data Mining Works with examples for Social Scientists in Higher Education Research", *CAIR*, Sacramento, November 2001.

[4] Hueglin C., Vannotti F., "Data Mining Techniques to Improve Forecast Accuracy in Airline Business", *In Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* ACM Press 2001, 438-442.

[5] Lawrence R.D., Hong S.J., Cherrier J., "Passenger-Based Predictive Modeling of Airline Now-show Rates", *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,*ACM Press 2003, Pages 397-406.

[6] Neapolitan R.E., *Learning Bayesian Network*s, Prentice Hall, 2004

[7] Jensen F., *Bayesian Networks and Decision Graphs*, Springer-Verlag, 2002.

[8] Witten I.H, Frank E., *Data Mining: Practical Machine Learning Tools and Techniques*, Second Edition, Morgan Kaufmann Publishers, 2005.

[9] Chawla N.V., Japkowicz N., Kotcz A., "Editorial: Special Issue on Learning from Imbalanced Data Set", *ACM SIGKDD Explorations*, Volume 6 Issue 1, ACM Press, 2004, Pages 1-6.

[10] Batista G.E.A, Prati R.C., Monard M.C., "A Study of the Behavior of Several Methods for Balancing Manchine Learning Training Data", *ACM SIGKDD Explorations*, Volume 6 Issue 1, ACM Press, 2004, Pages 20-29.

[11] Phua C., Alahakoon D., Lee V., "Minority Report in Fraud Detection: Classification of Skewed Data", *ACM SIGKDD Explorations*, Volume 6 Issue 1, ACM Press, 2004, Pages 50-59.

[12] Chawla N.V., Bowyer K.W., Hall L.O., Philip Kegelmeyer W.P., "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research 16*, Morgan Kaufmann Publishers, 2002, Pages 341-378.

[13] Chawla N.V., Lazarevic A., Bowyer K.W., Hall L.O., "SMOTEBoost: Improving Prediction of the Minority Class in Boosting", *7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Dubrovnik, Croatia, 2003, Pages 107-119,

[14] Guo H., Viktor H.L., "Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach", *ACM SIGKDD Explorations*, Volume 6 Issue 1, ACM Press 2004, Pages 30-39.

[15] Chan P., Fan W., Prodromidis A. and Stolfo S., "Distributed Data Mining in Credit Card Fraud Detection", *IEEE Intelligence Systems*, 14, 1999, Pages 67-74.

[16] Freund Y., Schapire R.E., "Experiments with a New Boosting Algorithms", *Proceedings of 13th International Conference on Machine Learning,* 1996, Pages 148-156.

[17] Dietterich T.G., "Ensemble Methods in Machine Learning", *First International Workshop on Multiple Classifier Systems*, Pages 1-15.