Expert Systems with Applications xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition

Dech Thammasiri [a], Dursun Delen [b,*], Phayung Meesad [c], Nihat Kasap [d]

[a] Faculty of Information Technology, King Mongkut's University of Technology North Bangkok Bangsue, Bangkok 10800, Thailand
[b] Spears School of Business, Department of Management Science and Information Systems, Oklahoma State University, Tulsa, OK 74106, USA
[c] Faculty of Information Technology, King Mongkut's University of Technology North Bangkok Bangsue, Bangkok 10800, Thailand
[d] School of Management, Sabanci University, Istanbul 34956, Turkey

ARTICLE INFO

Keywords:
Student retention
Attrition
Prediction
Imbalanced class distribution
SMOTE
Sampling
Sensitivity analysis

ABSTRACT

Predicting student attrition is an intriguing yet challenging problem for any academic institution. Class-imbalanced data is a common in the field of student retention, mainly because a lot of students register but fewer students drop out. Classification techniques for imbalanced dataset can yield deceivingly high prediction accuracy where the overall predictive accuracy is usually driven by the majority class at the expense of having very poor performance on the crucial minority class. In this study, we compared different data balancing techniques to improve the predictive accuracy in minority class while maintaining satisfactory overall classification performance. Specifically, we tested three balancing techniques—over-sampling, under-sampling and synthetic minority over-sampling (SMOTE)—along with four popular classification methods—logistic regression, decision trees, neuron networks and support vector machines. We used a large and feature rich institutional student data (between the years 2005 and 2011) to assess the efficacy of both balancing techniques as well as prediction methods. The results indicated that the support vector machine combined with SMOTE data-balancing technique achieved the best classification performance with a 90.24% overall accuracy on the 10-fold holdout sample. All three data-balancing techniques improved the prediction accuracy for the minority class. Applying sensitivity analyses on developed models, we also identified the most important variables for accurate prediction of student attrition. Application of these models has the potential to accurately predict at-risk students and help reduce student dropout rates.

1. Introduction

Increasing the student retention is a long term goal of any university in the US and around the world. The negative effects of student attrition are evident to students, parents, university and the society as a whole. The positive impact of increased retention is also obvious: college graduates are more likely to have a better career and have higher standard of life. College rankings, federal funding agencies, state appropriation committees and program accreditation agencies are all interested in student retention rates. Higher the retention rate, more likely for the institution to be ranked higher, secure more federal funds, traded favorably for appropriation and have easier path to program accreditations. Because of all of these reasons, administrator in higher education administrators are feeling increasingly more pressure to design and implement strategic initiatives to increase student retention rates. Furthermore, universities with high attrition rates face the significant loss of tuition, fees, and potential alumni contributions (Scott, Spielmans, & Julka, 2004). A significant portion of student attrition happens in the first year of college, also called the *freshmen year*. According to Delen (2011), fifty-percent or more of the student attrition can be attributed to the first year in the college. Therefore, it is essential to identify vulnerable students who are prone to dropping out in their freshmen year. Identification of the at-risk freshmen students can allow institutions to better and faster progress towards achieving their retention management goals.

Many modeling methods were found to assist institutions in predicting at-risk students, planning for interventions, to better understand and address fundamental issues causing student dropouts, and ultimately to increase the student retention rates. For many years, traditional statistical methods have been used to predict students' attrition and identify factors that correlate to their academic behavior. The statistics based methods that are more frequently used were logistic regression (Lin, Imbrie, & Reid, 2009; Scott et al., 2004; Zhang, Anderson, Ohland, & Thorndyke, 2004),

* Corresponding author. Tel.: +1 (918) 594 8283; fax: +1 (918) 594 8281.
E-mail address: dursun.delen@okstate.edu (D. Delen).
URL: http://spears.okstate.edu/~delen (D. Delen).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.046

discriminant analysis (Burtner, 2005) and structural equation modeling (SEM) (Li, Swaminathan, & Tang, 2009; Lin et al., 2009). Recently, many researchers have focused on machine learning and data mining techniques to study student retention phenomenon in higher education. Alkhasawneh (2011) proposed a hybrid model where he used artificial neural networks for performance modeling and used genetic algorithms for selecting feature subset in order to better predict the at-risk students and to obtain thorough understanding of the factors that relate to first year academic success and retention of students at Virginia Commonwealth University. Delen (2010) used a large and rich freshmen student data, along with several classification methods to predict attrition, and using sensitivity analysis, explained the factors that are contributing to the prediction models in a ranked order of importance. Yu, DiGangi, Jannasch-Pennell, Lo, and Kaprolet (2007) conducted a study where they used classification trees for predicting attrition and for identifying the most crucial factors contributing to retention. Zhang and Oussena (2010) proposed data mining as an enabler to improve student retention in higher education. The goal of their research was to identify potential problems as early as possible and to follow up with best possible intervention options to enhance student retention. They built and tested several classification algorithms, including Naïve Bayes, Decision Trees and Support Vector Machines. Their results showed that Naïve Bayes archived the highest prediction accuracy while the Decision Tree with lowest one.

This brief review of the previous studies shows that data mining methods have a great potential to augment the traditional means to better manage student retention. Compared to the traditional statistical methods, they have fewer restrictions (e.g., normality, independence, collinearity, etc.) and are capable of producing better prediction accuracies. Particularly when working with large data sets that contain many predictor variables, data mining methods proven to be robust in dealing with missing data, capturing highly complex nonlinear patterns, and hence producing models with very high level of prediction accuracy. Although, there is a consensus on the use of data mining and machine learning techniques, there is hardly any consensus on which data mining technique to use for the retention prediction problem. Literature has shown superiority of different techniques over the other in variety of different institutional settings. Depending on the data, and the formulation of the problem, any data mining technique can come out to be superior to any other. This lack of consensus prompts an experimental approach to identifying and using the most appropriate data mining technique for a given prediction problem. Therefore, in this study we developed and compared four different data mining techniques.

In the retention datasets, there usually are relatively fewer instances of students who have dropped out compared to the instances of students who have persisted. This data characteristic where the number of examples of one flaw type (i.e., a class label) is much higher than the others is known as the problem of imbalanced data, or the class imbalance problem. We found that in our dataset, minority class samples constitutedonly about 21% of thecomplete dataset. According to Li and Sun (2012) if the proportion of minority class samples constitutes less than 35% of the dataset, the dataset is considered as imbalanced. Therefore, in this study we are to deal with an imbalanced class distribution problem. The class imbalance problem is not unique to student retention, it is an intrinsic characteristics of many domains including credit scoring (Brown & Mues, 2012), prediction of liquefaction potential (Yazdi, Kalantary, & Yazdi, 2012), bankruptcy prediction (Olson, Delen, & Meng, 2012) and biomedical document classification (Laza, Pavon, Reboiro-Jato, & Fedz-Riverola, 2011). It has been reported in data mining research that when learning from imbalanced data, data mining algorithms tend to produce high predictive accuracy over the majority class, but poor predictive accuracy over the minority class. Learning from imbalanced data thus becomes an important sub field in data mining research. To improve the accuracy of classification methods with imbalanced data, several methods have been previously studied. These methods could be considered as a data preprocessing that take place before applying the classification methods. The methods to balance imbalanced data sets employ some variant of under sampling and/or over sampling of the original data sets.

In this research study, we developed and tested numerous prediction models using different sampling strategies such as under-sampling, over-sampling and SMOTE to handle imbalanced data. Using four different modeling techniques—logistic regression, decision tree, neural networks and support vector machines—over four different data structures—original, balanced with over-sampling, balanced with under-sampling and balanced with SMOTE—we wanted to understand the interrelationships among sampling methods, classifiers and performance measures to predict student retention data. In order to minimize the sampling bias in splitting the data between training and testing for each model building exercise, we utilized 10-fold cross validation. Overall, we executed a $4 \times 4 \times 10$ experimental design that resulted in 160 unique classification models. The rest of the paper is organized as follows: Section 2 provides a condensed literature review on student retention and the class imbalance problem. Section 3 describes the freshmen student dataset, and provides a brief review of the classification models, imbalance data techniques and evaluation metrics used for our study. Section 4 presents and discusses the empirical results. Section 5, the final section, concludes the paper with the listing of the contributions and limitations of this study.

2. Literature review

In this section, we first review the student retention problem from theoretical perspective—concept and theoretical models of student retention—and then review it from analytic perspective where machine learning and data mining techniques are used for classification of student attrition. In the second part of the section, we reviewed the literature on the methods used for handling class imbalance problem.

2.1. Student retention

There are two types of outcomes in student retention: *typical stayer* is a student enrolled each semester until graduation and graduates in due course plan; *a dropout*, or leaver, is a student who enters university but leaves prematurely or drop out before graduation and never returns to study again. High rates of student attrition have been reported in the reality of college readiness 2012 (see act.org).

Over the last several decades, researchers have developed the most comprehensive models (theoretical as well as analytic) to address higher education student retention problem. Earlier studies dealt with understanding the reasons behind student attrition by developing theoretical models. Undoubtedly the most famous researcher in this area is Tinto (1987). His student engagement model has served as the foundation for hundreds of other theoretical studies. Later, in addition to understanding the underlying reasons, the researchers have been interested in identifying at-risk students as early as possible so that they can prevent the likelihood of dropping out. Early identification of the students with higher risk of dropping out provides the means for the administrators to instigate intervention programs, provide assistance for those students in need. In earlier analytical approaches, traditional statistical methods such as logistic regression, discriminant analysis and

Please cite this article in press as: Thammasiri, D., et al. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications* (2013), http://dx.doi.org/10.1016/j.eswa.2013.07.046

structural equation modeling (SEM) were used most frequently in retention studies to identify factors and their contributions to the student dropout. Glynn, Sauer, and Miller (2002) developed the logistic regression model to provide early identification of freshmen at risk of attrition. The early identification is accomplished literally within a couple of weeks after freshman orientation. The model and its results were presented along with a brief description to the institutional intervention program designed to enhance student persistence. Luna (2000) used logistic regression, discriminant analysis, and classification and regression trees (CART). Focusing on new, incoming freshmen, this study examined several variables to see which can provide information about retention and academic outcome after three semesters. Scott et al. (2004) conducted a study which used a multiple linear regression to examine potential psychosocial predictors of freshman academic achievement and retention. This study demonstrated the utility of model to predict academic achievement but not college student retention. They suggested that future research should consider other psychosocial factors that might predict freshman retention. Pyke and Sheridan (1993) proposed a logistic regression analysis to predict the retention of master's and doctoral candidates at a Canadian university. Results for master's students indicate that analytic-driven interventions significantly improved the student's chances of graduating with the degree.

Some of the student retention research focused on students' academic experience and its derivatives. For instance, Schreiner (2009) conducted a research where he empirically linked student satisfaction to retention, postulating on the widespread belief that there is indeed a positive relationship between the two. His models focused on determining whether student satisfaction is predictive of retention the following year. A logistic regression analysis was conducted on each class level separately, using actual enrollment status as the dependent variable. Their results indicated that the first-year students whom do not find college enjoyable are 60% less likely to return as sophomores; while those with the sense of lack of belonging are 39% less likely, and those with difficulty of contacting their advisor are 17% less likely to return. On a related study, Garton and Ball (2002) conducted a research study to determine predictors of academic performance and retention of freshmen in the College of Agriculture, Food and Natural Resources (CAFNR) at the University of Missouri. Using the step-wise discriminant analysis method they built predictive models to determine whether a linear combination of student experience along with learning style, ACT score, high school class rank, and high school core GPA could determine the likelihood of persistence. In their linear models, high school core GPA was the best predictor of college academic performance for freshmen students. Furthermore, learning style was not a significant predictor of students' academic performance during their first year of enrollment in the college of agriculture. The traditional criteria used for college admission was found to have limited value in predicting agriculture students' retention.

Recently, machine learning and data mining techniques are gaining popularity in modeling and predicting student attrition. Yu, DiGangi, Jannasch-Pennell, and Kaprolet (2010) shown that their research attempts brought in a new perspective by exploring this issue with the use of three data mining techniques, namely, classification trees, multivariate adaptive regression splines (MARS), and neural networks; resulting in relatively better prediction models that identified transferred hours, residency, and ethnicity as crucial factors in determining student attrition. Lin (2012) also used a data mining approach to build predictive models for student retention management. His models aimed at identifying students who are in need of support from the student retention program using a variety of prediction models. The results show that some of the machine learning algorithms were able to establish reasonably good predictive models from the existing student retention data. Nandeshwar, Menzies, and Nelson (2011) studied to use data mining to find patterns of student retention at American Universities. They applied various attributes selection methods including CFS, Information Gain, chi-squared, and One-R to identify the ranked order importance of the independent variables. The researchers tested various classifiers such as One-R, C4.5, ADTrees, Naive Bayes, Bayes networks, and radial bias networks to create models for predicting student retention. Data used in this study were from a mid-size public university. After determining the subset of the attributes that best predict for student retention, the researchers conducted a contrast a set of experiments to seek attributes (values and ranges) that are most discriminative in various outcomes. They found that the rankings of all attribute ranges which, in isolation, predict for third year retention at a probability higher than the ZeroR limit (55%), and are supported by good number of records. The top six attributes most significant of third-year retention were the financial aid hypothesis: student's wages, parent's adjusted gross income, student's adjusted gross income, mother's in-come, father's income, and high school percentile.

In a more recent study, Lauría, Baron, Devireddy, Sundararaju, and Jayaprakash (2012) used a fall 2010 undergraduate students data from four different sources including students' biographic data and course related data; course management (Sakai—Sakai-Project.org) event data and Sakai's grade book data. They used oversampling to balance the data and applied three classifiers for prediction and comparison purposes. The models included logistic regression, support vector machines, and C4.5 decision trees. The result show that the logistic regression and the SVM algorithms provide higher classification accuracy than the C4.5 decision tree in terms of their ability to detect students at academic risk. Some of the other noteworthy recent studies in this domain include Kovačić, 2012; Yadav and Pal (2012), Yadav, Bharadwaj, and Pal (2012) among others. They all have used limited data sets with a wide range of machine learning techniques, finding somewhat different sets of predictors as the most important indicators of retention. Table 1 provides a tabular representation of some of the recent student retention studies and their data balancing and prediction model specifications. As the table clearly indicates, the class imbalance problem is not explicitly addressed in these studies (i.e., either not perceived as a significant problem or is not clearly explained in the article). The next sub-section provides a more detail about the class imbalance problem and its impact of the validity and value of prediction models.

2.2. The class imbalance problem

Imbalanced data problem is quite usual in machine learning and data mining applications as it appears in many real-world prediction tasks. However, the techniques and concept of balancing the data prior to model building is relatively new to many information systems researchers. A wide variety of balancing techniques have been applied to data sets in many areas such as medical diagnosis (Li, Liu, & Hu, 2010; Su, Chen, & Yih 2006), classifiers for database marketing (Duman, Ekinci, & Tanriverdi, 2012), property refinance prediction (Gong & Huang, 2012), classification of weld flaws (Liao, 2008) among others. In the class imbalance problems, the "imbalance" can be described as the number of instances in at least one class significantly outnumbering the other classes. We call the classes having more of the number of samples as the majority classes and the ones having fewer the number of samples as the minority classes. In such case, standard classifier algorithms usually have a bias towards the majority class (Xu & Chow, 2006; Zhou & Liu, 2006). These cases are shown in Fig. 1, where in the balanced data the accuracy over the minority class significantly in-

Please cite this article in press as: Thammasiri, D., et al. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications* (2013), http://dx.doi.org/10.1016/j.eswa.2013.07.046

Table 1
Comparisons of their applications that related work.

Work	Data balancing technique	Classification techniques
Yu et al. (2010)	–	Classification trees, Multivariate adaptive regression splines (MARS), and Neural networks (NN)
Lindsey, Lewis, Pashler, and Mozer (2010)	–	Percentage classifier, histogram classifier, logistic regression and BACT-R
Luna (2000)	–	Logistic regression, discriminant analysis and classification and regression trees (CART)
Garton and Ball (2002)	–	Step-wise discriminant analysis
Kovačić (2012)	–	Classification Trees, CHAID, CART and Logistic Regression
Yadav and Pal (2012)	–	C4.5, ID3 and CART decision tree algorithms
David and Renea (2008)	–	Logistic regression model
Yadav et al. (2012)	–	ID3, C4.5 and CART
Lin (2012)	–	ADT Tree, NB Tree, CART, J48 graft and J48
Nandeshwar et al. (2011)	–	One-R, C4.5, ADTrees, Naive bayes, bayes networks, and radial bias networks
Lauría et al. (2012)	Over-sampling method	Logistic regression (LR), Support vector machines (SVM) and C4.5 Decision trees
Zhang and Oussena (2010)	–	Naïve bayes, decision tree and Support vector machine
Alkhasawneh (2011)	–	Neural networks
Garton and Ball (2002)	–	Step-wise discriminant analysis
Lin et al. (2009)	–	Neural networks (NN), Logistic regression (LR), Discriminant analysis (DA) and Structural equation modeling (SEM)
Yu et al. (2007)	–	Classification trees
Salazar, Gosalbez, Bosch, Miralles, and Vergara (2004)	–	A decision rule based on C4.5 algorithm
Zhang et al. (2004)	–	Logistic regression
Li et al. (2009)	–	Logistic regression, stepwise/ hierarchical multiple regression, longitudinal data analysis, covariate adjustment, two-step design, exploratory factor analysis, classification tree, discriminant analysis and structural equation modeling
Herzog (2006)	–	Logistic regression, decision tree and neural networks
Veenstra, Dey, and Herrin (2009)	–	Logistic regression
Murtaugh, Burns, and Schuster (1999)	–	Multiple-variable model
Cabrera, Nora, and Castafne (1993)	–	Structural equations modeling (SEM)

creases while the accuracy over the majority one slightly decreases. Class imbalance usually degrades the real performance of a classification algorithm, by poorly predicting the minority class which is often the center of attention for a classification problem.

Many researchers have focused on the class imbalance problem in order to improve the prediction accuracy over the minority class. Some of these methods used a pre-processing approach, where they tried to balance the data before the model building, while others developed prediction algorithms that assign different weighting schemas to even-out class representations. The pre-processing approach seem to be the more straight forward approach that has greater promise to overcome class imbalance problem. This approach uses various methods to either randomly oversample the minority class or randomly under-sample the majority class (or some combination of the two). Random oversampling aims to balance class populations through creating new samples (from minority class by random selection) and adding them to the training set. On the other hand random under-sampling aims to balance the class populations through removing data samples from the majority class, until the classes are approximately equally represented. Even though there is no overwhelming evidence, the performances of under-sampling technique are thought to outperform the over-sampling technique (Drummond & Holte, 2003).

Although data balancing techniques are known to improve prediction results over the original data set, they have several important drawbacks. Namely, random over-sampling technique increases the size of the data set and therefore amplifies the computational burden. Random under-sampling may lead to lots of important information when examples of the majority class are randomly discarded from the original data set. Our brief review on recent research studies has revealed that the coverage on data balancing techniques is gaining more attention. There have been many research attempts to develop techniques to better balance the imbalance datasets prior to developing prediction models; these techniques are often derived from either over- or under-sampling some approached, such as SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), Borderline-SMOTE (Han, Wang, & Mao, 2005), Cluster-Based sampling (Taeho & Nathalie, 2004), Adaptive Synthetic Sampling algorithms (Haibo, Yang, Garcia, & Shutao, 2008), SMOTEBoost (Chawla, Lazarevic, Hall, & Bowyer, 2003), DataBoost-IM (Guo & Viktor, 2004).

3. Methodology

In this study, four popular classification methods—artificial neural networks, support vector machines, decision trees and logistic regression—along with three balancing techniques—random over-sampling, random under-sampling and SMOTE—are used to build prediction models, and compared to each other using 10-fold cross validation hold-out samples. As a result, in this study, we built 16 different types of classification models (each containing 10 experimental models), which are named and listed in Table 2. The classification performance measures are calculated using a 10-fold cross validation methodology. In this experimentation methodology the dataset is first partitioned into 10 roughly equal-sized distinct subsets. For each experiment nine subsets are used for training and the one part is used for testing. This procedure is repeated for 10 times for each of the 16 model types. Then test results are aggregates to portray the "unbiased" estimate of the model's performance. As shown in Eq. (1), the performance measure (PM) is averaged over k-folds (in this experimentation we set the value of k to 10). In the Eq. (1), CV stands for cross-validation, k is the number of folds used, and PM is the performance measure for each fold (Olson & Delen, 2008)

$$CV = \frac{1}{k}\sum_{i=1}^{k} PM_i \tag{1}$$

In order to demonstrate and validate the proposed methodology, two most popular data mining toolkit are used—IBMSPSS Modeler 14.2 and Weka 3.6.8. Fig. 2 shows an overview of our methodology (i.e., data preparation, model building and testing process).

3.1. Data

In this study we used seven years of institutional data (acquired from several disjoint databases), where the students enrolled as

Please cite this article in press as: Thammasiri, D., et al. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications* (2013), http://dx.doi.org/10.1016/j.eswa.2013.07.046

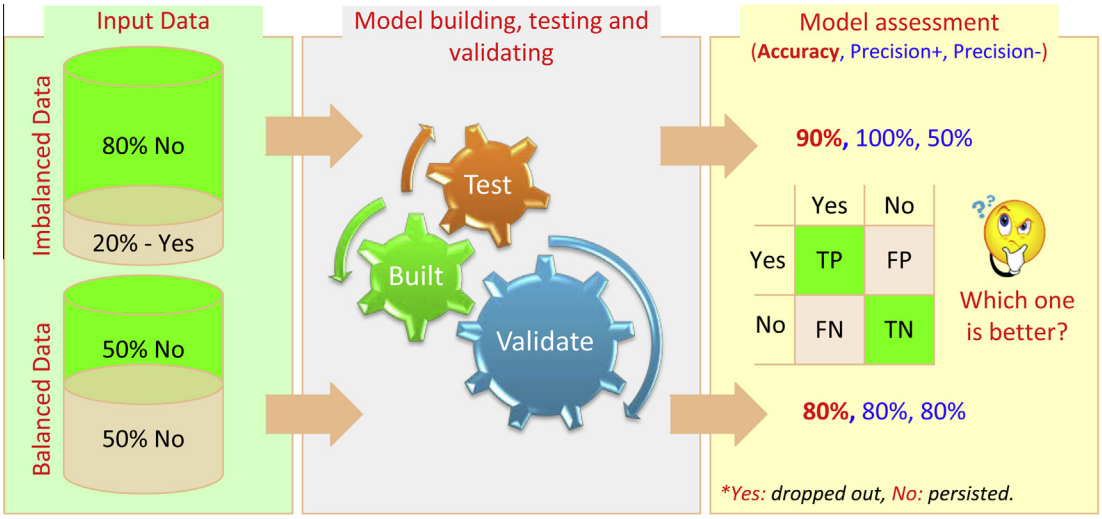

Fig. 1. The problem of imbalanced class distribution.

Table 2
Comparative models.

No.	Name	Description
1	LROR	Logistic regression with original data
2	DTOR	Decision tree with original data
3	ANNOR	Artificial neuron network with original data
4	SVMOR	Support vector machine with original data
5	LROS	Logistic regression with random over-sampling
6	DTOS	Decision tree with random over-sampling
7	ANNOS	Artificial neuron network with random over-sampling
8	SVMOS	Support vector machine with random over-sampling
9	LRUS	Logistic regression with random under-sampling
10	DTUS	Decision Tree with random under-sampling
11	ANNUS	Artificial neuron network with random under-sampling
12	SVMUS	Support vector machine with random under-sampling
13	LRSMOTE	Logistic regression with SMOTE
14	DTSMOTE	Decision tree with SMOTE
15	ANNSMOTE	Artificial neuron network with SMOTE
16	SVMSMOTE	Support vector machine with SMOTE

freshmen between (and including) the year 2005–2011. The dataset consisted of 34 variables and 21,654 examples/records, of which 17,050 were positive/retained (78.7%) and 4,604 were negative/dropped-out (21.3%). A brief summary of the number of records by year is given in Table 3. We performed a rigorous data preprocessing to handle anomalies, unexplainable outliers and missing values. For instance, we removed all of the international student records from the freshmen dataset (which was less than 4% of the total dataset) because they did not contain some of the presumed important predictors (e.g., high school GPA, SAT scores, among others).

The data contained variables related to student's academic, financial, and demographic characteristics. A complete list of variables obtained from the student databases is given in Table 4.

3.2. Classification methods

This study aims to compare the performance of four popular classification techniques within the student retention context. What follows is a short description of these four classification techniques.

3.2.1. Artificial neuron networks

Artificial neural network (ANN) is a computationally-intensive algorithmic procedure that transforms inputs into desired outputs using highly inter-connected networks of relatively simple processing elements (often called neurons, units or nodes). Neural networks are modeled after the neural activity in the human brain. Different network structures are proposed over the last few decades. For classification type problems (as is the case in this study), the most commonly used structure is called multi-layered perceptron (MLP). In MLP the network architecture consists of three layers of neurons (input, hidden and output) connected by weights, where the input of each neuron is the weighted sum of the network inputs, and the output of the neuron is a function (sigmoid or linear) value based on its inputs.

3.2.2. Support vector machines

Support vector machines (SVM) belong to a family of generalized linear models which achieves a classification model based on the linear combination of independent variables. The mapping function in SVM can be either a classification function (as is the case in this study) or a regression function. For classification, nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. Then, the maximum-margin hyperplanes are constructed to optimally separate the classes in the training data. The assumption is that the larger the margin or distance between these hyperplanes the better the generalization performance of the classifier. SVM are gain in popularity of being an excellent alternative to ANNs for prediction type problems.

3.2.3. Decision trees

Decision trees (DT) aim to predict discrete-valued target functions, where the learned function that connects the predictor variables to the predicted variable is represented by a decision tree (Mitchell, 1977). Decision tree algorithm uses a divide-and-concur methodology to find most discriminating variables and variable-values to create a tree-looking structure that is composed of nodes and edges. The main difference between different DT algorithms is the heuristic (Gini Index, Information Gain, Entropy, Chi-square, etc.) that they used to identify the most discriminating variable and variable-values. In this study, we used a popular decision tree algorithm (C5), which is an improved version of ID# and C4.5 algorithms developed by Quinlan (1986).

Please cite this article in press as: Thammasiri, D., et al. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications* (2013), http://dx.doi.org/10.1016/j.eswa.2013.07.046

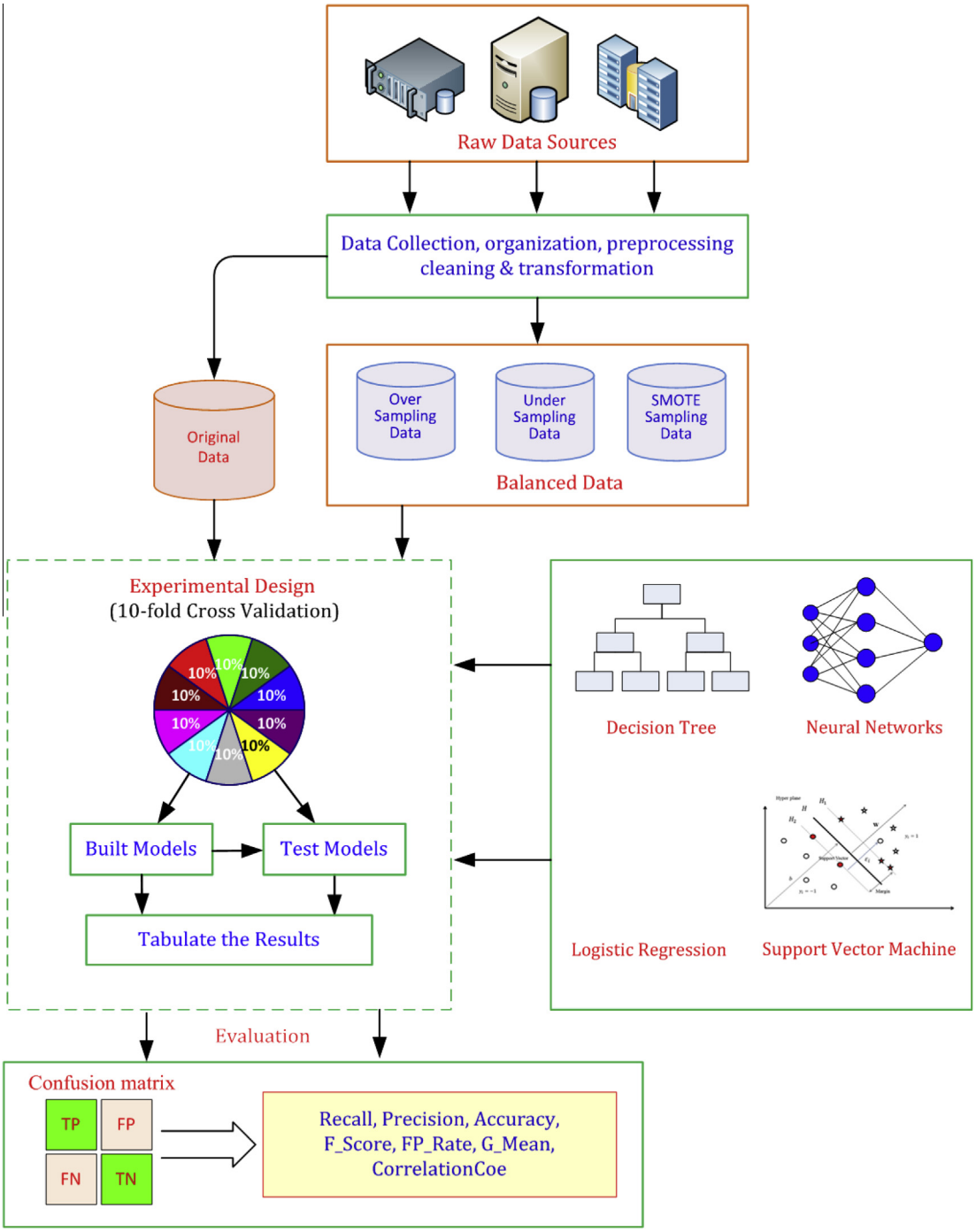

Fig. 2. The overview of the methodology employed in this study.

Table 3
Five-year freshmen student data.

Year	persisted freshmen	Dropped-out freshmen	Total students
2005	2419 (79.1%)	641 (20.9%)	3060
2006	2383 (80.0%)	595 (20.0%)	2978
2007	2290 (76.3%)	710 (23.7%)	3000
2008	2264 (78.9%)	605 (21.1%)	2869
2009	2284 (78.5%)	627 (21.5%)	2911
2010	2604 (79.4%)	674 (20.6%)	3278
2011	2806 (78.9%)	752 (21.1%)	3558
Total	17,050 (78.7%)	4604 (21.3%)	21,654

3.2.4. Logistic regression

Logistic regression is perhaps the most widely used classification technique, which has its roots in traditional statistics. The concept of the logistic regression is to examine the linear relationship between the dependent variables and independent variable. The dependent variable may be binomial (as is the case in this study) or multinomial.

3.3. Data sampling techniques

Sampling strategies are often used to overcome the class imbalance problem, whereby one of the two main approaches are pur-

Please cite this article in press as: Thammasiri, D., et al. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications* (2013), http://dx.doi.org/10.1016/j.eswa.2013.07.046

Table 4
Summary of data fields for the freshmen student data.

No.	Feature description	Type of data	Mean	Median	Std. deviation
1	College	Nominal	3.581007	3	1.888476
2	Degree	Nominal	7.611453	8	2.323752
3	Major	Nominal	1.596675	2	0.490582
4	Concentration specified	Nominal	1.248837	1	0.432354
5	Ethnicity	Nominal	7.134921	8	1.895381
6	Sex	Binary nominal	1.474206	1	0.499351
7	Residential code	Binary nominal	1.836549	2	0.369789
8	Marital status	Binary nominal	1.996853	2	0.056014
9	Admission type	Multi nominal	1.097222	1	0.388233
10	Permanent address state	Multi nominal	2.879105	3	0.410383
11	Received fall financial aid	Binary nominal	1.158046	1	0.364797
12	Received spring financial aid	Binary nominal	1.209428	1	0.406914
13	Fall student loan	Binary nominal	1.597906	2	0.490337
14	Fall grant/tuition waiver/scholarship	Binary nominal	1.224548	1	0.417299
15	Fall federal work study	Binary nominal	1.958744	2	0.198889
16	Spring student loan	Binary nominal	1.625479	2	0.484016
17	Spring grant/tuition waiver/scholarship	Binary nominal	1.274631	1	0.446343
18	Spring federal work study	Binary nominal	1.950876	2	0.216135
19	Fall hours registered	Number	14.38328	14	1.695436
20	Fall earned hours	Number	12.427	13	3.705255
21	Earned by registered	Number	0.862594	1	0.242983
22	Fall GPA	Number	2.772712	3	0.981879
23	Fall cumulative GPA	Number	2.825021	3.04	0.951662
24	SAT high score comprehensive	Number	24.0572	24	3.823869
25	SAT high score english	Number	23.96675	24	4.717876
26	SAT high score reading	Number	24.96367	25	5.048948
27	SAT high score math	Number	22.99822	23	4.474751
28	SAT high score science	Number	23.62308	23	3.950641
29	Age	Number	18.51704	18	0.661422
30	High school GPA	Number	3.536636	3.6	0.383027
31	Years after high school	Number	0.039135	0	0.41521
32	Transferred hours	Number	1.916325	0	4.680315
33	CLEP hours	Number	0.748495	0	3.335043
34	Second fall registered	Binary nominal	0.577176	1	0.816648

sued—ether eliminating some data from the majority class (under-sampling) or adding some artificially generated or duplicated data to the minority class(over-sampling).

3.3.1. Random under-sampling (RUS)

This is a non-heuristic method that randomly selects examples from the majority class for removal without replacement until the remaining number of examples is roughly the same as that of the minority class.

3.3.2. Random over-sampling (RUS)

This method randomly select examples from the minority class with replacement until the number of selected examples plus the original examples of the minority class is roughly equal to that of the majority class.

Table 5
A typical confusion matrix for a binary classification problem.

		Predicted results	
		Predicted positive	Predicted negative
Actual Results	Actual positive	True positives (TP)	False positives (FP)
	Actual negative	False negatives (FN)	True negative (TN)

3.3.3. Synthetic minority over-sampling technique (SMOTE)

This heuristic, originally developed by Chawla et al. (2002), generates synthetic minority examples to be added to the original dataset. For each minority example, its k nearest neighbors of the same class is found. Some of these nearest neighbors are then randomly selected according to the over-sampling rate. A new synthetic example is generated along the line between the minority example and every one of its selected nearest neighbors. This process is repeated until the number of examples in all classes is roughly equal to each other.

3.4. Evaluation measures

To evaluate the performance of 16 classification methods, we used a number of popular metrics. These metrics are calculated using the confusion matrixes (see Table 5). Confusion matrix is a unique tabulation of correctly and incorrectly predicted examples for each class. For a binary classification problem, there are four populated cells: True Positives (TP)—denote the number of positive examples that were predicted correctly, True Negatives (TN)—denote the number of negative examples that were predicted correctly, False Positives (FP)—denote the number of positive examples that were predicted incorrectly, and False Negatives (FN)—denote the number of negatives examples that were predicted incorrectly.

Following are the performance measures used in evaluating and comparing prediction models

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$specificity = \frac{TN}{TN + FP} \tag{4}$$

$$precision^{+} = \frac{TP}{TP + FP} \tag{5}$$

$$precision^{-} = \frac{TN}{TN + FN} \tag{6}$$

$$FP\text{-}Rate = \frac{FP}{FP + TN} \tag{7}$$

$$F\text{-}measure = \frac{2^{*}Precision^{*}Recall}{Precision + Recall} \tag{8}$$

$$CC = \frac{TP^{*}TN - FN^{*}FP}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \tag{9}$$

$$GMEAN = \sqrt{sensitive^{*}specificity} \tag{10}$$

Please cite this article in press as: Thammasiri, D., et al. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications* (2013), http://dx.doi.org/10.1016/j.eswa.2013.07.046

4. Experimental results

The results of all 16 models on all 9 performance measures are listed in Table 6. Each cell is populated with mean and standard deviation of the respective performance measure. Fig. 3 presents the average accuracy, sensitivity and specificity of the classification models. On average, SVMSMOTE model provided the highest rate of accuracy (0.902) and specificity (0.958). DTOS model provides the highest rate of sensitivity (0.885). In addition, the methods that using on balance data are still provide greater specificity that the methods that using on unbalance data.

In this study, the ranked importance of the predictor factors was also investigated to discover the relative contribution of each to the prediction models. In order to understand the relative importance of the features used in the study, we conducted a sensitivity analysis on trained prediction models, where we measured the comparative importance of the input features in predicting the output. Once the importance factors determined for each of the 16 models, they are aggregated and combined for ranking purposes. Table 7 shows the top 10 predictor variables. As can be seen, the most important factors came out to be FallGPA, HrsEarned/Registered, SpringGrantTuitionWaiverScholarship, ReceivedSpringAid, SpringStudentLoan, SATHigh- Science, SATHighEnglish, Ethnicity, FallStudentLoan, MajorDeclared.

5. Discussion and conclusion

In this paper, we described our approach to the construction of classifiers from imbalanced datasets. A dataset is imbalanced if the classes (i.e., classification categories) are not nearly equally represented. Most real-world data sets are imbalanced, containing a large number of regular/expected examples with only a small percentage of irregular/unexpected examples. Often, what is interesting is the recognition/prediction of the irregular/unexpected examples. Machine learning techniques are not very good at discerning/predicting less representative class in imbalanced datasets. Therefore a data balancing task is needed as part of the data preprocessing phase. In this study, we compared three data balancing techniques using four popular classification methods along with a large feature-rich real-world data set.

To succeed, college student retention projects should follow a multi-phased process, which may starts with identifying, storing (in a databases), and using student data/characteristics to better understand underlying reason and to predict the at-risk students who are more likely to dropout, and ends with developing effective and efficient intervention methods to retain them. In such a process, analytics can play the most crucial role of accurately identifying students with the highest propensity to drop-out as well as explaining the factors underlying the phenomenon. Because machine learning methods (such as the ones used in this study) are capable of modeling highly nonlinear relationships, they are believed to be more appropriate techniques to predict the complex nature of student attrition with a high level of accuracy.

The results of this study show that, if proper methods of preprocessing applied to sufficiently large data sets with the rich set of variables, analytics methods are capable of predicting freshmen student attrition with high level of accuracy (as high as 90%). SMOTE balancing technique combined with support vector machine classification method provided the highest overall performance (i.e., prediction accuracy, correlation coefficient and G-mean). From the usability standpoint, despite the fact that SVM and ANN had better prediction results, one might chose to use decision trees because compared to SVM and ANN, they portray a more transparent model structure. Decision trees explicitly show the reasoning process of different prediction outcomes, providing a justification for a specific prediction, whereas SVM and ANN are mathematical models that do not provide such a transparent view of how they do what they do.

A noteworthy strength of this study is that it provides a rank-ordered importance of the features used in the perdition modeling. Specifically, sensitivity analysis is applied to prediction models to identify their comparative importance (i.e., additive contribution) in predicting the output variable. The sensitivity values of all variables across all 16 model types are aggregated to construct the final list of variable-value pairs. Such an understanding not only help build more parsimonious models, but also helps decision makers understand what variables are the most important in improving retention rates.

The success of analytics project relies heavily on the richness (quantity and dimensionality) of the data representing the phenomenon being considered. Even though this study used a large sample of data (covering several years of freshmen student records) with a rather rich set of features, more data and more variables can potentially help improve the analytics/prediction results. Some of the variables that have a great potential to improve prediction performance include student's social interaction/connectedness (being a member of a fraternity or other social groups); student's parent's or significant others educational and financial backgrounds, and student's prior expectation/ambitions from his educational endeavors.

Potential future directions of this study include (i) extending the predictive modeling methods to include ensembles (model combining/fusing techniques), (ii) enhancing the information

Table 6
Ten-fold cross validation classification performance measures for all models.

Model	Accuracy	Sensitivity	Specificity	Precision+	Precision-	FP-Rate	F-Measure	Corr. Coef.	G-Mean
LROR	0.864 (±0.006)	0.874 (±0.006)	0.794 (±0.017)	0.966 (±0.003)	0.485 (±0.028)	0.794 (±0.017)	0.918 (±0.003)	0.549 (±0.023)	0.833 (±0.010)
DTOR	0.864 (±0.008)	0.877 (±0.008)	0.783 (±0.040)	0.962 (±0.011)	0.502 (±0.039)	0.783 (±0.040)	0.918 (±0.005)	0.553 (±0.029)	0.829 (±0.021)
ANNOR	0.860 (±0.005)	0.878 (±0.006)	0.754 (±0.020)	0.955 (±0.005)	0.507 (±0.027)	0.754 (±0.020)	0.915 (±0.003)	0.540 (±0.021)	0.814 (±0.011)
SVMOR	0.864 (±0.007)	0.867 (±0.007)	0.840 (±0.020)	**0.977** (±0.003)	0.444 (±0.035)	0.840 (±0.020)	**0.919** (±0.004)	0.545 (±0.029)	0.853 (±0.012)
LROS	0.774 (±0.005)	0.736 (±0.005)	0.827 (±0.010)	0.855 (±0.011)	0.693 (±0.010)	0.827 (±0.010)	0.791 (±0.005)	0.556 (±0.011)	0.780 (±0.005)
DTOS	0.844 (±0.007)	**0.885** (±0.023)	0.812 (±0.010)	0.793 (±0.018)	0.896 (±0.024)	0.812 (±0.010)	0.836 (±0.007)	0.693 (±0.017)	0.848 (±0.009)
ANNOS	0.771 (±0.004)	0.733 (±0.006)	0.826 (±0.009)	0.855 (±0.011)	0.687 (±0.011)	0.826 (±0.009)	0.789 (±0.004)	0.551 (±0.009)	0.778 (±0.005)
SVMOS	0.785 (±0.006)	0.745 (±0.006)	0.842 (±0.010)	0.869 (±0.010)	0.702 (±0.011)	0.842 (±0.010)	0.802 (±0.006)	0.579 (±0.012)	0.792 (±0.006)
LRUS	0.775 (±0.001)	0.738 (±0.014)	0.828 (±0.017)	0.860 (±0.017)	0.688 (±0.025)	0.828 (±0.017)	0.794 (±0.011)	0.557 (±0.025)	0.782 (±0.012)
DTUS	0.770 (±0.001)	0.729 (±0.013)	0.832 (±0.019)	0.867 (±0.020)	0.671 (±0.025)	0.832 (±0.019)	0.792 (±0.011)	0.549 (±0.024)	0.779 (±0.012)
ANNUS	0.768 (±0.014)	0.735 (±0.014)	0.815 (±0.020)	0.847 (±0.020)	0.688 (±0.022)	0.815 (±0.020)	0.787 (±0.013)	0.542 (±0.027)	0.774 (±0.014)
SVMUS	0.779 (±0.016)	0.736 (±0.017)	0.846 (±0.017)	0.879 (±0.015)	0.677 (±0.028)	0.846 (±0.017)	0.801 (±0.013)	0.569 (±0.031)	0.789 (±0.015)
LRSMOTE	0.801 (±0.004)	0.753 (±0.007)	0.849 (±0.005)	0.832 (±0.008)	0.775 (±0.009)	0.849 (±0.005)	0.790 (±0.004)	0.604 (±0.008)	0.799 (±0.004)
DTSMOTE	0.896 (±0.010)	0.856 (±0.012)	0.934 (±0.009)	0.925 (±0.011)	**0.871** (±0.012)	0.934 (±0.009)	0.889 (±0.010)	0.793 (±0.019)	0.894 (±0.010)
ANNSMOTE	0.854 (±0.012)	0.812 (±0.015)	0.895 (±0.014)	0.881 (±0.018)	0.832 (±0.015)	0.895 (±0.014)	0.845 (±0.013)	0.710 (±0.025)	0.852 (±0.013)
SVMSMOTE	0.902 (±0.004)	0.849 (±0.008)	**0.958** (±0.004)	0.954 (±0.004)	0.860 (±0.009)	**0.958** (±0.004)	0.898 (±0.004)	**0.810** (±0.007)	**0.902** (±0.004)

Please cite this article in press as: Thammasiri, D., et al. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications* (2013), http://dx.doi.org/10.1016/j.eswa.2013.07.046

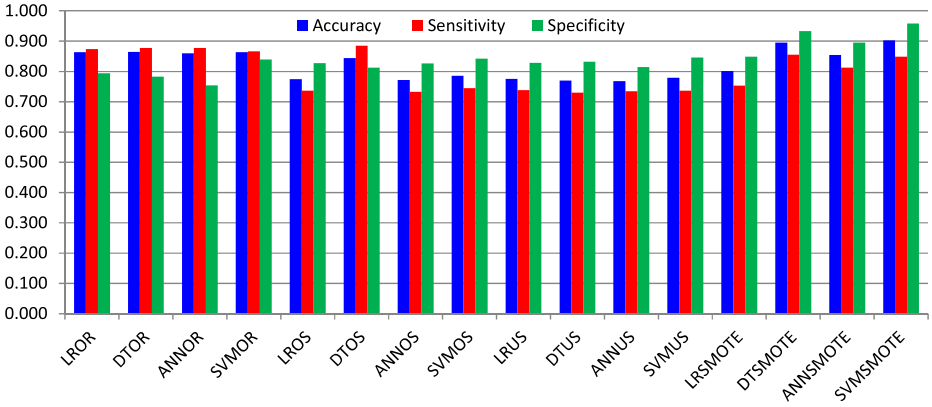

Fig. 3. Average accuracy, sensitivity and specificity of the classification models.

Table 7
Ten most important features with feature importance ranking in DTOR model.

Rank	Feature	Feature importance score
1	FallGPA	0.602
2	HrsEarned/Registered	0.492
3	SpringGrantTuitionWaiverScholarship	0.461
4	ReceivedSpringAid	0.236
5	SpringStudentLoan	0.229
6	SATHighScience	0.125
7	SATHighEnglish	0.118
8	Ethnicity	0.087
9	FallStudentLoan	0.076
10	MajorDeclared	0.046

sources by including the data from survey-based institutional studies (which are intentionally crafted and carefully administered for retention purposes), and perhaps most importantly, (iii) deployment of the information system as a decision aid for administrators, so that the pros and cons of the systems would be assessed for improvement and better fit to the institutional needs.

References

Alkhasawneh, R. (2011). *Developing a hybrid model to predict student first year retention and academic success in STEM disciplines using neural networks*. Virginia Commonwealth University.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446–3453.

Burtner, J. (2005). The use of discriminant analysis to investigate the influence of non-cognitive factors on engineering school persistence. *Journal of Engineering Education, 94*, 335–338.

Cabrera, A. F., Nora, J.-E. A., & Castafne, M. B. (1993). Structural equations modeling test of an integrated model of student retention. *The Journal of Higher Education, 64*(2), 123–139.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 341–378.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In *Seventh European conference on principles and practice of knowledge discovery in databases (PKDD)* (pp. 107–119). Dubrovnik, Croatia.

David, F., & Renea, F. (2008). Predictors of first-year student retention in the community college. *Community College Review, 36*(2), 68–88.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498–506.

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention, 13*(1), 17–35.

Drummond, C., & Holte, R. C. (2003). C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II, ICML*. Washington, DC.

Duman, E., Ekinci, Y., & Tanriverdi, A. (2012). Comparing alternative classifiers for database marketing: The case of imbalanced datasets. *Expert Systems with Applications, 39*, 48–53.

Garton, B. L., & Ball, A. L. (2002). The academic performance and retention of college of agriculture students. *Journal of Agricultural Education, 43*(1), 46–56.

Glynn, J. G., Sauer, P. L., & Miller, T. E. (2002). A logistic regression model for the enhancement of student retention: The identification of at-risk freshmen. *International Business & Economics Research Journal, 1*(8), 79–86.

Gong, R., & Huang, S. H. (2012). A Kolmogorov–Smirnov statistic based segmentation approach to learning from imbalanced datasets: With application in property refinance prediction. *Expert Systems with Applications, 39*, 6192–6200.

Guo, H., & Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach. *SIGKDD Explorations Newsletter, 6*, 30–39.

Haibo, H., Yang, B., Garcia, E. A., & Shutao, L. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the IEEE international joint conference on neural networks, IJCNN 2008. (IEEE World Congress on Computational Intelligence)* (pp. 1322–1328).

Han, H., Wang, W. -Y., & Mao, B. -H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *First international conference on intelligent computing (ICIC)* (pp. 878–887). Hefei, China.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research, 131*, 17–33. John Wiley & Sons, Inc. Published online in Wiley Interscience.

Kovačić, Z. J. (2012). Predicting student success by mining enrolment data. *Research in Higher Education Journal, 15*, 1–20.

Lauría, E. J. M., Baron, J. D., Devireddy, M., Sundararaju, V., & Jayaprakash, S. M. (2012). Mining academic data to improve college student retention: An open source perspective. In *Second international conference on learning analytics and knowledge* (pp. 139–142) Vancouver, BC, Canada.

Laza, R., Pavon, R., Reboiro-Jato, M., & Fedz-Riverola, F. (2011). Evaluating the effect of unbalanced data in biomedical document classification. *Journal of Integrative Bioinformatics, 8*(3), 1–13.

Li, D.-C., Liu, C.-W., & Hu, S. C. (2010). A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine, 40*, 509–518.

Li, H., & Sun, J. (2012). Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples – evidence from the Chinese hotel industry. *Tourism Management, 33*(3), 622–634.

Li, Q., Swaminathan, H., & Tang, J. (2009). Development of a classification system for engineering student characteristics affecting college enrollment and retention. *Journal of Engineering Education, 98*, 361–376.

Liao, T. W. (2008). Classification of weld flaws with imbalanced class data. *Expert Systems with Applications, 35*, 1041–1052.

Lin, S. H. (2012). Data mining for student retention management. *Journal of Computing Sciences in Colleges, 27*(4), 92–99.

Lin, J. J. J., Imbrie, P. K., & Reid, K. J. (2009). Student retention modelling: An evaluation of different methods and their impact on prediction results. In *Proceedings of the research in engineering education symposium 2009* Palm Cove, QLD.

Lindsey, R., Lewis, O., Pashler, H., & Mozer, M. (2010). Predicting students' retention of facts from feedback during study. In *Proceedings of the 32nd annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

Luna, J. (2000). Predicting student retention and academic success at new mexico tech, New Mexico Institute of Mining and Technology, Socorro, New Mexico.

Mitchell, T. M. (1977). *Machine learning*. New York: McGraw-Hill.

Murtaugh, P. A., Burns, L. D., & Schuster, J. (1999). Predicting the retention of university students. *Research in Higher Education, 40*(3), 355–371.

Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications, 38*(12), 14984–14996.

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Berlin, Heidelberg: Springer-verlag.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining models for bankruptcy prediction. *Decision Support Systems, 52*(2), 464–473.

Please cite this article in press as: Thammasiri, D., et al. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications* (2013), http://dx.doi.org/10.1016/j.eswa.2013.07.046

Pyke, S. W., & Sheridan, P. M. (1993). Logistic regression analysis of graduate student retention. *Canadian Journal of Higher Education, 23*(2), 44–64.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.

Salazar, A., Gosalbez, J., Bosch, I., Miralles, R., & Vergara, L. (2004).A case study of knowledge discovery on academic achievement, student desertion and student retention. In *Second international conference on information technology: Research and education, 2004, ITRE 2004* (pp. 150–154).

Schreiner, L. A. (2009). Linking student satisfaction and retention. Research study: Azusa Pacific University.

Scott, D. M., Spielmans, G. I., & Julka, D. C. (2004). Predictors of academic achievement and retention among College freshmen: A longitudinal study. *College Student Journal, 38*(1), 66–80.

Su, C.-T., Chen, L.-S., & Yih, Y. (2006). Knowledge acquisition through information granulation for imbalanced data. *Expert Systems with Applications, 31*, 531–541.

Taeho, J., & Nathalie, J. (2004). Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter, 6*, 40–49.

Tinto, V. (1987). Leaving college: Rethinking the causes and cures of student attrition. University of Chicago Press, 5801 S. Ellis Avenue, Chicago, IL 60637.

Veenstra, C. P., Dey, E. L., & Herrin, G. D. (2009). A model for freshman engineering retention. *Advances in Engineering Education, 1*(3), 1–33.

Xu, L., & Chow, M.-Y. (2006). A classification approach for power distribution systems fault cause identification. *IEEE Transactions on Power Systems, 21*(1), 53–60.

Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. *International Journal of Innovative Technology & Creative Engineering, 1*(12), 13–19.

Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. *World of Computer Science and Information Technology Journal (WCSIT), 2*(2), 51–56.

Yazdi, J. S., Kalantary, F., & Yazdi, H. S. (2012). Prediction of liquefaction potential based on CPT up-sampling. *Computers & Geosciences, 44*(0), 10–23.

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science, 8*, 307–325.

Yu, F. H., DiGangi, E., Jannasch-Pennell, A., Lo, W., & Kaprolet, C. (2007). A data-mining approach to differentiate predictors of retention. Educause Southwest Conference. Austin, TX.

Zhang, Y., & Oussena, S. (2010). Use data mining to improve student retention in higher education - a case study. In *12th International conerence on enterprise information systems*. Portugal.

Zhang, G., Anderson, T. J., Ohland, M. W., & Thorndyke, B. R. (2004). Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. *Journal of Engineering Education, 93*(4), 313–320.

Zhou, Z.-H., & Liu, X.-Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering, 18*, 63–77.

Please cite this article in press as: Thammasiri, D., et al. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications* (2013), http://dx.doi.org/10.1016/j.eswa.2013.07.046