Profiling Waitlisted Incoming Students for Future Delinquency with an Ensemble of Statistical Machine Learning Algorithms*

Maureen Lyndel C. Lauron and Jaderick P. Pabico

Institute of Computer Science

University of the Philippines Los Baños

{mclauron, jppabico}@up.edu.ph

Abstract

Given a dataset $\mathcal{R} = \{R_1, R_2, \ldots, R_r\}$ of r records of waitlisted incoming freshman students (WIFS), where for any $i = 1, 2, \ldots, r$, R_i is a $(m+1)$tuple $(O_i, P_i^{(1)}, P_i^{(2)}, \ldots, P_i^{(m)})$, O_i is any one in a set $\mathcal{O} = \{O_1, O_2, \ldots, O_o\}$ of o classes, and $P_i^{(1)}, P_i^{(2)}, \ldots, P_i^{(m)}$ are m potential predictors for O_i. Our purpose is to find a statistical machine learning algorithm (SMLA) $\mathbb{A}$ such that $V_i = \mathbb{A}(P_i^{(1)}, P_i^{(2)}, \ldots, P_i^{(m)})$, where V_i is a predicted class by $\mathbb{A}$ that was developed using $n \leq m$ correct number of predictors for $O \in \mathcal{O}$, and $\mathbb{A}$ is the best algorithm such that the metric $v^{-1}\sum_{i=1}^{v}|O_i - V_i|$ is minimum across $v < r$ records in the validation set $\mathcal{V} \subset \mathcal{R}$. Our problem is to find the subset $\{P_i^{(1)}, P_i^{(2)}, \ldots, P_i^{(n)}\}$ and to train $\mathbb{A}$ using $t < r$ records from the training set $\mathcal{T} \subset \mathcal{R}$, such that $\mathcal{T} \cap \mathcal{V} = \emptyset$, so that $\mathbb{A}$ can predict whether a WIFS trying to enter an undergraduate program at UPLB will incur at least a "delinquency" once the student is accepted into the program. The $\mathbb{A}$ can be a useful decision-support tool for UPLB deans and college secretaries in deciding whether a WIFS will be accepted into the program or not.

The potential predictors $P_i^{(1)}, P_i^{(2)}, \ldots, P_i^{(m)}$ are the ith WIFS own UPCAT record such as gender, age, high school grade, province, UPCAT score, etc. In this problem, $m = 21$. The set $\mathcal{O}$ is composed of $o = 5$ classes, the first four of which are considered by the administration as "delinquencies." These classes are: (1) The student will transfer to another UP campus after being accepted into a program (O_1); (2) The student will incur poor scholastic performance in the program (O_2); (3) The student will shift to a different program (O_3); (4) The student will commit absence without leave or file for leave of absence or honorable dismissal (O_4); and (5) The student will continue with the program (O_5). The desirable predicted class using any $\mathbb{A}$ should be O_5 where the decision for student acceptance into the program becomes a trivial one.

Based on UPLBs freshmen intake record of AYs 2011-2012, 2012-2013, and 2013-2014 furnished by the Office of the University Registrar (OUR), $r = 2,302$. The dataset, however, is heavily inbalanced in favor of O1 comprising about 59% of $\mathcal{R}$, which means that every 3 of 5 WIFS chose to transfer to another UP campus after having been accepted into the program, seemingly using the program as a stepping stone to the campus that the WIFS did not qualify to. The rest are 5%, 2%, 1%, and 33% for O_2,

*Submitted as a scientific oral paper contribution to the 18th National Student-Faculty Conference on the Statistical Sciences, SEARCA, Los Baños, Laguna, 16 October 2017.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3312v1 | CC BY 4.0 Open Access | rec: 3 Oct 2017, publ: 3 Oct 2017

O_3, O_4, and O_5, respectively. With this skew, any $\mathbb{A}$ will just have to classify a WIFS as either a O_1 or a O_5 for a 92% classification accuracy. Thus, we needed to implement a class cardinality balancing method $\mathbb{B}^*$ over $\mathcal{R}$ from among a set $\mathbb{B} = \{B_1, B_2, \ldots, B_b\}$ of b available methods, so that $|\{O_1\}| \approx |\{O_2\}| \approx \cdots \approx |\{O_5\}|$. Thus, applying any B_j over $\mathcal{R}$ will result to $r_j \neq r$ for any $j = 1, 2, \ldots, b$. After applying the Synthetic Minority Oversampling Technique (SMOTE), for example, $r_j = 3,110$.

We have to choose the best algorithm $\mathbb{A}^*$ from a bigger set $\mathcal{A}^{[5]} = \mathcal{A} \times \mathcal{A}^3 \times \mathcal{A}^5$, where $\mathcal{A} = \{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_a\}$ is a list of available SMLAs, $\mathcal{A}^3 = \mathcal{A} \times \mathcal{A} \times \mathcal{A}$ is an ensemble of three algorithms chosen from $\mathcal{A}$, and $\mathcal{A}^5 = \mathcal{A} \times \mathcal{A} \times \mathcal{A} \times \mathcal{A} \times \mathcal{A}$ is an ensemble of five. Examples of $\mathcal{A}$ are the family of k-nearest neighbors, support vector machines, artificial neural networks, C4.5 decision trees, Bayesian networks, etc. In this work, $a = 37$. This means that the search space for $\mathbb{A}^*$ is $a + a^3 + a^5$, which can be "exhaustively" searched via a step-wise forward substitution procedure. Similarly, we have to select the proper dimension reduction technique $\mathbb{D}^*$ from among a set $\mathcal{D} = \{D_1, D_2, \ldots, D_d\}$, examples of which are Principal Component Analysis (PCA) and Genetic Algorithm (GA).

Using a directed yet "exhaustive" search method from the combination $\mathbb{B} \times \mathbb{A}^{[5]} \times \mathbb{D}$, we found out that $\mathbb{B}^*$ is SMOTE, $\mathbb{A}^*$ is bagging ensemble seeded with C4.5 decision trees, and $\mathbb{D}^*$ is PCA providing 73%, 43%, 43%, 68%, and 96% respective prediction rates for classes O_1, O_2, O_3, O_4, and O_5, all based on a 5×5 confusion matrix of $\mathbb{A}^*$ over the records in $\mathcal{V}$. The overall prediction rate is 79.7%.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3312v1 | CC BY 4.0 Open Access | rec: 3 Oct 2017, publ: 3 Oct 2017