Profile-based Recruiting of New Students

Ray R. Hashemi[1], Louis A. Le Blanc[2], Azita Bahrami[3], Kevin Willett[1], and
Xaunna J. Krehn[1]

[1]Department of Computer Science
Armstrong Atlantic University
Savannah, GA 31419, USA
Ray.Hashemi@armstrong.edu

[2]Campbell School of Business
Berry College
Mount Berry, GA 30149-5024, USA
lleblanc@berry.edu

[3]IT Consultation Company
Savannah, GA, USA
Azita.G.Bahrami@gmail.com

Abstract – **A profile-based recruiting of new students for an institution of higher education is more efficient, financially sound, and more successful. In this paper, two different methodologies, Aprioi Algorithm and Modified Rough Sets, are used to create profiles from historical data collected by an admissions office of a college in the southeast United States. The first approach delivered two and the second approach delivered five profiling rules. The profiling rules were evaluated against a test set. The success rate of the Apriori Algorithm and Modified Rough Sets were 87% and 75%, respectively. The first approach had the false positive of 6% a false negative of 4%. The second approach had a false positive of 10% and false negative of 2%.**

***Keywords*–Profiling; Profiling Rules; Recruiting; Apriori Algorithm; Modified Rough Sets.**

I. INTRODUCTION

An institution of higher education receives many applications from prospective students. After a lengthy process of screening applications, a select number of new students are offered seats at the institution. Often students who have been offered these seats withdraw their applications and they do not show up for classes. These students have strong GPAs, good standardized tests scores, and therefore they receive multiple offers and then choose their favorite. As a result, the respective admissions offices accept more than their capacities and develop "alternate" or "wait" lists of applicants.

The money spent on recruiting, advertising, and long hours for screening of the applications consumes a sizable chunk of the admissions budget [1, 2]. These costs can be reduced substantially if a profile of potential or likely new students was known. That is, advertising will be tailored toward this targeted group of students and the advertisements will appear only in locations that reach potential students. The number of students who do not matriculate should drop, suggesting a more successful recruiting process.

The goal of this research is to generate profiles of those students who might attend the institution once accepted. Upon establishing such a profile, all the recruiting activities are channeled to those students who meet the profile.

The organization of the remainder of this paper is: relevant background in Section 2, the methodology in Section 3, empirical results in Section 4, and conclusion in Section 5.

II. RELEVANT BACKGROUND

Both approaches, the Apriori Algorithm and the Rough Sets, are introduced in the next two subsections, respectively. (The concept of the modified Rough Sets is discussed in sub-section 3.3.)

A. Apriori Algorithm

A record with n attributes consists of n predicates of $A_i(x_{ij})$ (for i = 1to n and j = 1 to m), where A_i is the i-th attribute with m possible values and x_{ij} is the j-th possible value for A_i. In reference to the goal of the study, a predicate's value represents one piece of information about a student who applies for a seat in a university. In addition, one predicate is designated as the decision predicate. Its value represents the admissions office's action on the student. The predicates other than the decision predicate are referred to as condition predicates. We use the Apriori Algorithm [3] to establish the association(s) between the decision predicate and the condition predicates. The algorithm identifies the predicate sets that appear together most frequently. If k predicates frequently appear together in a dataset, they make a k-dimensional predicate set (k-D predicate set).

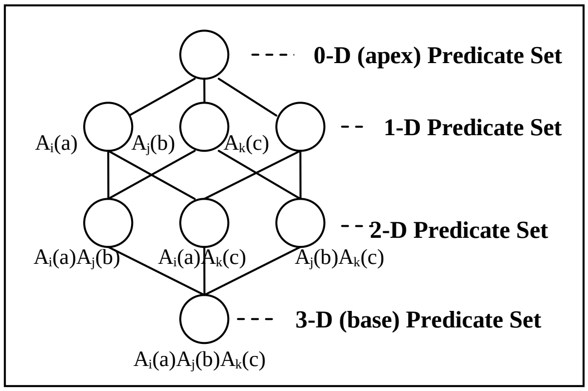

Figure 1. A predicates' lattice for a dataset with three most frequent predicates of $A_i(a)$, $A_j(b)$ and $A_k(c)$

The predicates of a predicate set reflect a stronger bond among participant predicates. The predicate set which includes a decision predicate is of interest because it shows a strong bond between condition predicates and decision predicate. These predicate sets are used to build association rules.

The algorithm starts by checking the frequency of appearance of each individual predicate in all records of a dataset to identify those predicates whose frequency is greater than a threshold (*support count*) t. The outcome makes 1-D predicate sets. As the second step, the most frequent 2-D predicate sets are identified. To do so, all 2 by 2 possible combinations of the 1-D predicate sets are built; and, those with support count less than threshold t are filtered. The process continues until the most frequent k-D predicate set is identified. The value for k is decided when (k+1)-D predicate set is empty because all of them are filtered out. All the frequent predicate sets for a dataset may be shown in form of a predicates' lattice, Fig. 1. The number of possible predicates for n attributes is :

$$\sum_{j=1}^{n}\frac{n!}{j!(n-j)!} \qquad (1)$$

For a relatively large n, the number of predicates is too many. The use of a support count threshold weeds out a large number of the predicates in each predicate set.

B. Rough Sets

The Rough Sets approach was introduced by Pawlack in 1984 [4]. The details can be found in [5, 6, 7]. First, the Rough Set approach is defined, and then it is put in perspective to the problem at hand.

Definition 1: An approximate space P is an ordered pair P(U, R), where U is the universe of objects and R is a binary equivalence relation over U.

Definition 2: Let R* be a family of subsets of R and let $A \subseteq U$. If for some $Y \subseteq R^*$, A is equal to the union of all sets in Y, then A is definable in P; otherwise, A is non-definable or A is *a rough set*.

Definition 3: Any Rough Set A has a lower approximation space Low(A), an upper approximation space Up(A), and a boundary B(A). And they are defined as follows:

$$\mathrm{Low}(A) = \{a \subseteq U \mid [a]_R \subseteq A], \qquad (2)$$

$$\mathrm{Up}(A) = \{a \subseteq U \mid [a]_R \cap A \neq \emptyset], \qquad (3)$$

$$B(A) = \mathrm{Up}(A) - \mathrm{Low}(A). \qquad (4)$$

Let the relatively large rectangle, Fig. 2, represent a dataset (universe), and each small rectangle represent a student record. Small rectangles of the same shade are records with the same condition predicates. There are four such sets in Fig. 2, namely S1 (all black rectangles), S2 (all gray rectangles), S3 (all rectangles with pattern), and S4 (all the white rectangles). Let the possible values for a decision predict be m and one of these values is x_a. The records that have the same decision predicate of decision(x_a) are shown bordered by the broken line and make a Rough Set, because it cannot be created by any possible union of the four sets. The lower approximation of the Rough Set includes those four sets that are totally inside the rough set (i.e., S1). The upper approximation of the Rough Set includes those four sets that are totally or partially inside the Rough Set (i.e., S1, S2, and S3). The boundary of the Rough Set includes S2 and S3. The condition predicates belonging to the lower approximation of a decision predicate have a stronger bond with the decision value than those in the upper approximation space.

Since there are more than one value for the decision predicate, there are more than one Rough Set for the dataset. Thus, the methodology is named Rough Sets (plural).

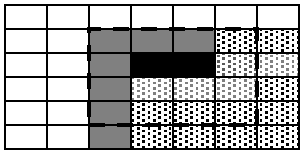

Figure 2. Rough sets visualization

III. METHODOLOGY

To create profiles of new students who are highly likely to attend the university after they are accepted, we use the Apriori Algorithm and the Modified Rough Sets approach separately for building the profiles. The profiling process is completed using the following steps:

1. Cleaning historical data.
2. Discovery of association rules from the cleaned historical data.
3. Discovery of approximation rules from the cleaned historical data using Modified Rough Sets approach.
4. Building the profiles.

The resulting profiles are evaluated on a sample set obtained from the original dataset to compare the performance of the two approaches.

A. Cleaning Historical Data

Historical data was collected by an admissions office of a college in the southeast United States. The historical data is cleaned vertically and horizontally. The vertical reduction is done by removing (a) the duplicate records (objects) from the historical dataset and (b) records with missing data.

The horizontal reduction is done by removing the redundant attributes from the vertically reduced dataset. The entropy approach [3] is used to identify the redundant attributes. To explain further: Let one or a set of attributes of the dataset be the *decision attributes* and the rest of them be *condition attributes*. In addition, let a decision attribute have m distinct values (classes). (In the case that the decision attribute is made up of more than one attribute, i.e. complex attribute, the classes for the complex decision attribute are all the possible combinations of the classes of the constituents.)

To determine the redundant condition attributes:

1. The entropy of the set of condition attributes, C, is calculated as follows:

$$E(C) = -\sum_{i=1}^{m} p_i * log_2 * p_i \qquad (5)$$

Where, p_i is frequency of class i in the dataset.

2. For each condition attribute q in C which has $v_1, v_2, . . ., v_n$ possible values, the *information gain* is calculated using formulas 6 and 7.

$$B(q) = \sum_{i=1}^{n} N_{v_i} * E(v_i) \qquad (6)$$

Where, N_{v_i} is the number of records in the dataset with $q = v_i$ and $E(v_i)$ is the entropy of these records.

$$\text{Gain}(q) = E(c) - B(q) \qquad (7)$$

3. If Gain(q) is less than a chosen threshold value, the attribute q is considered redundant and it is removed from the dataset.

B. Discovery of Strong Association Rules

Consider a record of a given dataset. This record is composed of a set of condition predicates and decision predicates. A predicate is composed of an attribute and its value. The condition predicates are considered to be the conditions under which a process takes place. The decision predicates are considered to be the outcome of the process.

For example, the record in Fig. 3 represents an application of a student who has applied for admission along with the respective decisions made by the institution and applicant. The first five attributes make the condition predicates and attributes, Accepted and Matriculated make the decision predicates. To explain further, Accepted(1) and Accepted(0) mean the student was or was not admitted; and Matriculated(1) and Matriculated(0) mean the student enrolled or did not enroll at the university.

We apply the Apriori Algorithm to the dataset and obtain all the frequent k-predicate sets. A k-predicate set is composed of k predicates. Let k_1-predicate sets and k_2-predicate sets be two subsets of a k-predicate set such that $k_1 \cup k_2 = k$ (condition 1) and $k_1 \cap k_2 = \emptyset$ (condition 2). The association rules of $k_1 \Rightarrow k_2$ and $k_2 \Rightarrow k_1$ are generated out of k-predicates.

There are several subsets of k that satisfy the conditions (1) and (2), therefore, several association rules are generated from one frequent k-predicate set.

To each associate rule two measurements of *support* and *confidence* are assigned using the following formulas:

$$\text{Support } (k1 \Rightarrow k2) = P(k1 \cup k2)/M, \qquad (8)$$

$$\text{Conf } (k1 \Rightarrow k2) = P(k1 \cup k2)/P(k1). \qquad (9)$$

For the above formulas, $P(k_1 \cup k_2)$ and $P(k_1)$ are the number of records with $(k_1 \cup k_2)$ and (k_1) in the dataset, respectively. M = |dataset|.

ID	Zip Code	GPA	SAT	Income	Accepted	Matriculated

Figure 3. A record layout

Filtering of the association rules in form of $k_i \Rightarrow k_j$ are done based on the following set of principles:

Principle 1: Rules with confidence less than a selected threshold are pruned.
Principle 2: If $k_i \cap$ (Decision attributes) $\neq \varnothing$, then rule is pruned.
Principle 3: If $k_j \cap$ (Condition attributes) $\neq \varnothing$, then rule is pruned.

Principle 1 ensures that the higher the confidence, the stronger the rule. Principles 2 and 3 deliver the *inter-dimension* association rules for the decision attributes.

C. *Modified Rough Sets*

In Rough Set nomenclature, an information system, S, is a quadruple (U, Q, V, d)
Where,
U is a non-empty finite set of objects, u.
Q is a finite set of attributes, q.
V $= \cup_{q \in Q}$Vq, and Vq is the domain of attribute q.
d is a mapping function such that d (a,q) $\in$ Vq for every q $\in$ Q and a $\in$ U.

All the objects who have the same values for their condition attributes constitute one *class*. And all the objects who have the same values for their decision attributes constitute one *partition*. The number of partitions is equal to the number of decision values. Consider partition λ_i, and classes $c_1, \ldots, c_n$. The objects of those classes that are totally contained within p_1 make the *lower approximation space* of $\lambda_{i,}$ Low(λ_i). The objects of all the classes that are either "totally" or "partially" contained in λ_i make the *upper approximation space* of λ_i, Up(λ_i). The objects of all the classes that are "partially" contained in p_1 make the *boundary* of λ_i, B(λ_i). The objects in boundary of λ_i have the same set of condition attributes but different decisions.

In any statistical model, these objects are removed because they are conflicting. However, we use the Modified Rough Sets approach to salvage the conflicting objects. Conflicting objects are part of life. For example, two patients (objects) with the same symptoms may be diagnosed differently (conflicting objects). Or, two prospective students with the same set of conditions, one decides to attend the college and the other one does not.

One of the decision values in B(λ_i) is designated as the *dominant decision* using Bayes' Theorem [5]. The decision values for all the subjects in B(λ_i) are then changed to the dominant decision. Such modification makes B(λ_i) = $\varnothing$. And, therefore the Rough Sets are changed into Modified Rough Sets. The rules that are generated from the objects of a Modified Rough Set are called *approximate rules* [5]. Each approximate rule has a certainty factor that is the same as the probability assigned to its decision value by Bayes' theorem, Formula 10.

$$P[d_i|\mathrm{Class}_j)\,\frac{P[\mathrm{Class}_j\;|d_{i]}]}{\sum_{i=1}^{m} P[\mathrm{Class}_j\;|d_{i]}]*Q_j]} \qquad (10)$$

Where, d_i is the i-th decision value and Class_j is all the records with the same set of condition values.

In Modified Rough Sets, those classes of data in which all subjects have the same decision values, the dominant decision is the common decision value and the probability assigned to such a dominant decision is 1.

D. *Building the Profiles*

Profiles are meta-rules that are built from a set of rules. This is completed through a collapsing process that integrates and generalizes rules. The following three guidelines govern the collapsing process.

Guideline1:
The following rules of r1 and r2 are given:
r1: ATT_1=a1 $\wedge$ ATT_2 = b3 $\wedge$ ATT_3 = c4 $\wedge$ ATT_4 = d2 → ATT_d = f2
r2: ATT_1=a1 $\wedge$ ATT_2 = b3 $\wedge$ ATT_3 = c4 $\wedge$ ATT_4 = d2 $\wedge$ ATT_5 = e1 → ATT_d = f2
The rule r1 reads "If Attribute#1 is equal to a1 and Attribute#2 is equal to b3 and Attribute #3 is equal to c4 and Attribute#4 is equal to d2, then Attribute decision = f2".

All the objects that fire rule r2 are a subset of objects firing rule r1. Therefore, r1 and r2 are collapsed into a new rule that is the same as the rule r1.

Guideline 2:
The following rules of r1 and r2 are given:
r1: ATT_1=a1 $\wedge$ ATT_2 = b3 $\wedge$ ATT_3 = c4 $\wedge$ ATT_4 = d2 → ATT_d = f2
r2: ATT_1=a2 $\wedge$ ATT_2 = b3 $\wedge$ ATT_3 = c4 $\wedge$ ATT_4 = d2 → ATT_d = f2
The rules r1 and r3 may collapse into a new rule
r': ATT_1 = (a1$\vee$ a2) $\wedge$ ATT_2 = b3 $\wedge$ ATT_3 = c4 $\wedge$ ATT_4=d2 → ATT_d= f2
If a1, and a2 are the only possible values for attribute ATT_1, then r' changes into
r": ATT_2 = b3 $\wedge$ ATT_3 = c4 $\wedge$ ATT_4=d2 → ATT_d = f2

Guideline 3 (heuristic rule):
The following rules of r1 and r2 are given:
r1: ATT_1=a1 $\wedge$ ATT_2 = b3 $\wedge$ ATT_3 = c4 $\wedge$ ATT_4 = d2 → ATT_d = f2
r2: ATT_1= k3 $\wedge$ ATT_2 = b3 $\wedge$ ATT_3 = c4 $\wedge$ ATT_4 = d2 → ATT_d = f2
a = |r1.conditions| and
b = |r2.conditions|

If $|r1 \cap r2| > Thcommon \wedge$
$|r1 - r2| < Thdifference \wedge$
$| r2 - r1| < Thdifference$
Then r1 and r2 can be collapsed into a new rule that is the same as r1 (if $a \le b$) or the same as r5 (if $a > b$).
The Thcommon and Thdifference are two thresholds decided by the analyst.

IV. EMPIRICAL RESULTS

Three different JAVA programs were developed to implement data cleaning, Apriori Algorithm and Modified Rough Sets. All three programs were executed on an Hewlet Packard laptop.

The historical dataset had over 30,000 records and each record had 20 attributes. Only 1,577 records survived the vertical cleaning and ten attributes survived the horizontal cleaning. The ten attributes along with their values and meanings are described in Table 1.

TABLE I. NON-REDUNDANT ATTRIBUTES AND THEIR VALUES

Attribute	Values
Parent's Zip	1 (All zip codes belong to Atlanta), 2 (All GA zip codes of Atlanta Suburbs), and 3 (All other zip codes)
Gender	1(Female) and 2(Male)
Ethnicity	1 (Caucasian), 2 (Others)
State	1(Georgia), 2(Other States)
SAT Score	1 (≤ 1000), 2 (> 1000 and ≤ 1200), and 3 (> 1200)
GPA:	1 (<3.00), 2 (≥3.00 and <3.5), 3 (≥3.50 and <4.00), and 4 (≥4.00)
Family Contribution	1(= 0), 2 (>0 and <=5000), 3(>5000 and ≤ 15000), 4(>15000 and ≤ 25000), and 5 (>25000)
Accepted	0 (No), 1 (Yes)
Matriculated	0 (No), 1 (Yes)
Cancelled	0 (No), 1 (Yes)

The numbers of association rules obtained by applying Apriori Algorithm, their minimum and maximum confidence levels along with the number of profiling rules are shown in Table 2. The numbers of approximate rules generated by the Modified Rough Sets approach along with the profiling rules are displayed in Table 3.

To check the validity of the profiles, we have applied the set of profiles on the test set. The test set has 159 records (roughly 10% of the total records). The test results are shown in Table 4.

The profiling results, Table 4, using profile rules of the Apriori Algorithm has 87% correct profiling with false positive of 6% and false negative of 4%. Using the profile rules of the Modified Rough Sets approach has 75% correct profiling with false positive of 10% and false negative of 2%.

TABLE II. ASSOCIATION RULE STATISTICS

Decision Attribute	**Association Rules**			**No. Profiling Rules**
	No.	**Min Conf**	**Max Conf**	
Matriculated	43	75%	79%	1
Not Matriculated	8	68%	71%	1

Profile 1:
If Parent Zip = 3 ∧ SAT =3, Then Matriculation = 0 (Conf = 68%)
Profile 2:
If Parent Zip = 2 ∧ Gender = 1 ∧ SAT ≥ 2 ∧ Family contribution = 5, Then Matriculation = 1 (Conf= 75%)

TABLE III. APPROXIMATE RULE STATISTICS

Decision Attribute	**Approximate Rules**			**No. Profiling Rules**
	No.	**Min Conf**	**Max Conf**	
Matriculated	107	60%	100%	4
Not Matriculated	159	60%	100%	2

Profile 1:
If Parent Zip = 1 ∧ Family Contribution = 0 ∧ GPA>=4.00 ∧ SAT >1200
Then Matriculation =1 (100%)
Profile 2
If Parent Zip = 1 ∧ Family Contribution = 4 ∧ 3.00 <=GPA <=3.5
Then Matriculation =1 (60%)
Else Matriculation =0 (40%)
Profile 3:
If Parent Zip = 1 ∧ Family Contribution = 5 and GPA >=3.00
Then Matriculation =1 (100%)
Profile 4:
If Parent Zip = 3 ∧ Family Contribution = 5 and GPA <3.00,
Then Matriculation =1 (66%)
Profile 5:
If Parent Zip = 3 ∧Family Contribution = 0,
Then Matriculations = 0;

ISBN: 978-1-61208-237-0

TABLE IV. RESULTS

	No. Record Match the Profiles	**No. Record with Correct Profiling**	**No. Record with Incorrect Profiling**	**No. Record Match No Profiles**
Apriori Algorithm	152	138	14	7
Modified Rough Set	137	119	18	22

V. CONCLUSION

The results in Table 4 reveal that 81% and 75% of the records have been correctly profiled for the predicates of Decision(Matriculated) and Decision(Not Matriculated) by the Apriori Algorithm. For the Modified Rough Sets approach the success rate is 50% for Decision(Matriculated) and 100% for Decision (Not Matriculated). The results seem rather diverse between the two algorithms. The reason may stem from the fact that the Apriori Algorithm acts at the condition predicates level, whereas the Modified Rough Sets approach acts at the record level. In other words, the first approach builds a set of the most frequent predicates regardless of the concern about record boundaries, but the Modified Rough Sets build the approximate rules with rigid concern about the record boundaries.

For the current dataset hand, results show profile-based recruiting of new students may save time and money. The saving is accomplished by concentrating the recruiting efforts on specific geographical areas with potential students who will matriculate.

REFERENCES

[1] Sun N. and Z. Yang, ”Equilibria and indivisibilities: gross substitutes and complements”, Econometrica, 2004; 74-5: 1385-1402.

[2] Abizada A., “Pairwise stability and strategy-proofness for college admissions with budget constraints”, Social Science Electronic Publishing, Inc. 2012.

[3] Han J. and Kamber M, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[4] Pawlak Z. “Rough Classification”, *Journal of Man-Machine Studies* 1984; 20: 469-83.

[5] Hashemi R., Pearce B., Arani R., Hinson W., Paule M. “A Fusion of Rough Sets, Modified Rough Sets, & Genetic Algorithms for Hybrid Diagnostic Systems”, In: Lin TY, Cercone N Editors. *Rough Sets & Data Mining: Analysis of Imprecise Data*. Kluwer Academic Publishers, 1997. pp. 149-76.

[6] Hashemi R., Tyler A., Bahrami A., "Use of Rough Sets as a Data Mining Tool for Experimental Bio-Data”, In *Computational Intelligence in Biomedicine and Bioinformatics: Current Trends and Applications*, Tomasz G. Smolinski, Mariofanna G. Milanova, and Aboul Ella Hassanien, Editors, Springer-Verlag Publisher, June 2008, pp. 69-91.

[7] Hashemi R., Choobineh F., Slikker W., and Paule M., "A Rough-Fuzzy Classifier for Database Mining", The *International Journal of Smart Engineering System Design*, No. 4, 2002, pp.107-114.