

An Automated University Admission Recommender System for Secondary School Students

Simon Fong and Robert P. Biuk-Aghai

Abstract—University or college admission is a complex decision process that goes beyond simply matching test scores and admission requirements. Past research has suggested that students' backgrounds and other factors correlate to the performance of their tertiary education. However, almost all admission and enrollment studies are based on the perspective of universities or colleges, and only few studies are based on the perspective of secondary schools. This paper presents a hybrid model of neural network and decision tree classifier that serves as the core design for a university admission recommender system. The system was tested with live data from sources of Macau secondary school students. In addition to the high prediction accuracy rate, flexibility is an advantage such that the system can predict suitable universities that match the students' profiles and the suitable approaches through which the students should enter. The recommender can be generalized into making different kinds of predictions based on the students' histories.

Index Terms—university admission, recommender system, classification, neural network, decision tree.

I. INTRODUCTION

A university education has become a basic part of most people's preparation for working life. Admission to university is therefore a topic of importance. How a student chooses a university, and conversely how a university chooses a student, determines the success of both sides in carrying through the education.

However, most existing studies of university admission [1],[2],[3] are based on the perspective of universities who are to receive the new incoming students, and not on the perspective of secondary schools that are sending their students to pursue higher education, or on the perspective of the student who has to decide which university to apply to. Given that the university knows very little about the applicant, whereas the secondary school knows a great deal more, there is value in extending the university admission process to include secondary schools.

This paper proposes a novel design of a recommender system that can provide recommendations about which universities a student should apply to, taking not only the student's secondary school scores but also other factors into account. By combining both a decision tree approach and a

Both authors are members of the Business Intelligence Group at the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Taipa, Macau, China (email: ccfong@umac.mo, robertb@umac.mo).

© ICITA 2009 ISBN: 978-981-08-3029-8

neural network approach, an improved recommendation output can be achieved.

The remainder of this paper is organized as follows. Section II discusses the problem of student recommendation in more detail. Section III then presents the design of our recommender system, RSAU (Recommender System of Admission to University), and Section IV evaluates the performance of our system. Section V discusses the decision rules used by RSAU. Finally, Section VI makes conclusions.

II. RECOMMENDATION PROBLEM

The choice of a university that is suitable for a given secondary school graduate can be a difficult decision to make. Reputation of the university, perceived difficulty of the degree program, distance from home, tuition and living costs, student's areas of academic strength as well as actual scores achieved are just some of the factors that may be considered by a student graduating from secondary school. Likewise, the university has its own set of admission criteria, mainly based on academic standard of the student to be admitted, but possibly also including others, such as minority and gender representation, local vs. domestic vs. overseas student proportion, and others. Choosing the most suitable among the many thousands of candidates that apply to a university every year is not a trivial matter. Some universities avoid many of these issues through a simple unified admission process, such as pre-defined secondary school completion scores required for admission. This approach, however, does not always result in the most suitable candidates reaching the university "best" for them. Moreover, many accepted candidates end up not taking up the offer extended to them, resulting in wasted administrative effort on the part of the university. Most importantly, however, it assumes that secondary schools have comparable education standards and curricula. When this assumption is not valid then student selection for university admission becomes problematic. In this paper we focus on this particular problem, and show in the sections that follow our design of a recommender system suitable to countries and territories in which country-/territory-wide education standards are not defined. For illustration purposes and as a case study we examine this problem in detail for the case of Macau.

Macau is a small territory, formerly a colony of Portugal and since 1999 a Special Administrative Region of China. It has a population of about 0.5 million, mainly of Chinese background but also including about 5% of non-Chinese speaking nationals, including those for whom Portuguese, English, and others are the languages primarily spoken. Moreover, the Chinese

population can be further divided into those speaking Mandarin, Cantonese and other Chinese dialects, as well as those coming from mainland China who are accustomed to using the simplified Chinese script, and the local Chinese population who use the traditional Chinese script. As a result of this fragmentation, many different kinds of primary and secondary schools exist, differentiated mainly 1) by their medium of instruction: Chinese-only, Portuguese-only, English-only, mainly-Chinese partly-English, mainly-English partly-Chinese, Chinese and Portuguese in about equal proportions, Chinese and English in about equal proportions; 2) by script: simplified Chinese vs. traditional Chinese; 3) by curriculum: based on mainland China, Hong Kong, Singapore, Portugal, UK, USA, Canada, Australia, a mix of the above, or a local self-developed one; 4) by orientation: science vs. arts vs. business; 5) by approach: conservative Chinese (characterized by large classes, rote learning, memorization, and much homework) vs. progressive Western (characterized by small classes, encouragement of creative skills, individual student tutoring, and moderate amount of homework), to name some of the most obvious differences. The Education and Youth Affairs Bureau of the Macau Government imposes no requirements on the curriculum taught at any of the schools, only requiring that they be government-registered. Thus each school has full autonomy in deciding its own curriculum. This results in one of the most complex pre-tertiary education systems in the world. It is therefore no surprise that the level of students leaving Macau's secondary schools varies greatly from one to another.

Admitting these students to any of the ten local tertiary institutions is mainly decided through an entrance examination. As there is no single territory-wide higher education entrance examination, each institution designs and administers its own examinations, focusing mainly on English, Chinese, mathematics, and sciences. A second option, particular to Macau, is the so-called "direct admission" where secondary school teachers can nominate top students for entrance to university without having to take any admission examination. A small quota of students for direct admission is allocated and distributed proportionally among the secondary schools of Macau.

Beyond the borders of Macau there are many more choices, among which mainland China, Taiwan and Western countries including the USA, Canada, UK and Australia are popular. Each of these has its own admission requirements and may mandate its own entrance examination that students need to sit. Thus for a student who is unsure of which university to enter and considers applying to several the workload of preparing for and sitting multiple examinations is high. From the point of view of secondary school teachers, they face a complex decision-making problem of 1) filling the quota of direct admission without entrance examination by selecting top students; 2) finding the best allocation of universities for the selected top students; and 3) likewise for choosing the medium-quality students to have a chance to continue to university but who will need to sit an entrance examination. A simplified model of such traditional process of deciding how

the students should be recommended is shown in Fig. 1.

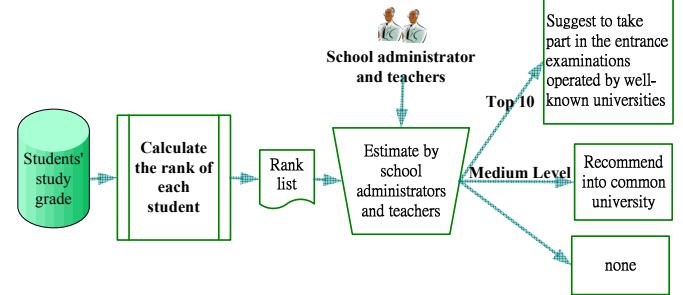


Fig. 1. Traditional method to make recommendation decision for secondary school graduates' admissions.

The traditional process suffers mainly from the inaccurate prediction of matching the right students to the right universities as feedback from the secondary school teachers, due to subjective human judgments. Moreover, it is not a single attribute estimation on which a decision is made. There exist other factors than academic scores that influence on university admissions. There is no single rule governing how a student should take up his/her tertiary course. It is a complex problem by nature with consideration of many attributes, for example, major of the course in arts or sciences, Gender (M or F), the student's background, etc. Therefore it is desirable to have some automated recommender system that take into account of multiple variables like demographic information, secondary school origin, major, study status, etc. for assisting secondary school teachers to make decision of recommendation. Figures 2a and 2b, show a conceptual model of a recommender which is able to make such predictions in an automated way, for the sake of decision support.

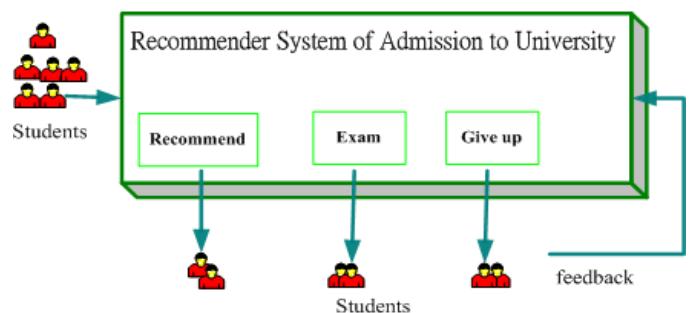


Fig. 2a. Classifying students to different groups by different university admissions ways.

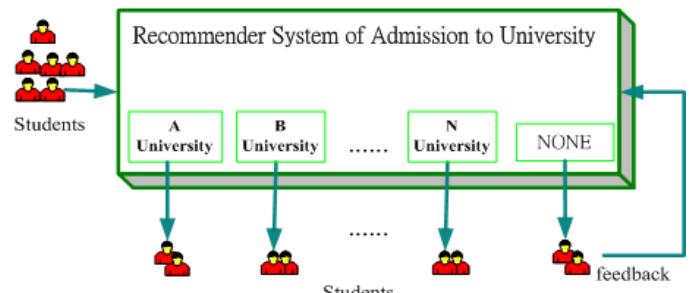


Fig. 2b. Analyzing various students data for predicting their university admissions.

III. RSAU RECOMMENDER SYSTEM

To the best of our knowledge, there are few recommender systems specifically built for analyzing pedagogical information for secondary school students by predicting their chances of university admission. Although technologically there may be numerous excellent data mining programs available, they are not tailored for generating recommendations for education sectors. The individual data mining tools are too technical and have very comprehensive features, making them difficult for general users.

Recommender System of Admission to University (RSAU) is specialized to analyze various data of secondary school students for predicting their university admissions, and classifying them to different groups by different university admissions ways (either by recommendation of direct-entry, by taking entrance exams or not going to any university). The end-users of RSAU system are secondary school administrators, teachers, course co-coordinators, policy makers who are involved in the admission process.

The architecture of RSAU has three main components, namely Data Analyzer, Classifier and Visualization that provides an easy-to-use interface for non-technical users. An overall diagram is shown in Fig. 3.

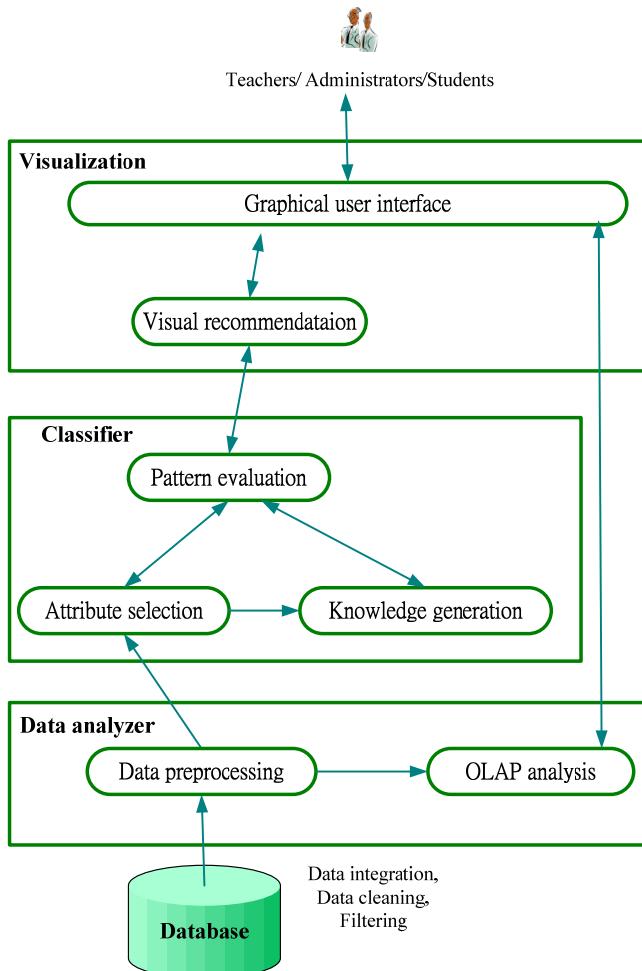


Fig. 3. Architecture of our proposed recommender system.

A. Data Analyzer

The Data Analyzer module mainly contains two major components, Data preprocessing and OLAP analysis. Their functions are as follow.

Data Preprocessing: to translate and reformat attributes of student records into suitable data types. Data preprocessing is important [4] because real-world data tend to be incomplete, noisy and inconsistent. The first step is “Data cleaning” which attempts to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. When the cleaned data is fed into “Data transformation” routine, the corresponding type of attributes is assigned. Another task is to divide the full dataset into two parts, one for learning and training, and another for testing and verification. The translated and reformatted data will be sent to the Attribute selection component in Classifier module.

OLAP analysis: to summarize, consolidate, view, apply formulae to, and synthesize data to multiple dimensions. It helps users to better understand the data, and browse them by any pairs of dimensions at will.

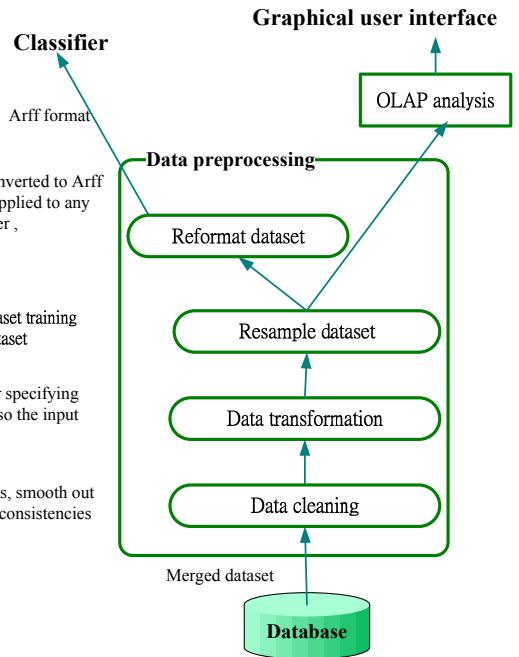


Fig. 4. The data processing flow of Data Analyzer.

B. Classifier

The Classifier module is the core of RSAU recommender system. The main function is classifying and predicting students' admission to universities based on students' data, which cumulated by their secondary school records, such as test grades and student profile data etc.

Robustness and stability are two important criteria for any prediction tool. In order to meet the criteria, we adopted a hybrid approach. Out of many variables that are used to characterize students' profiles, we need to efficiently find out those that have most impact on admissions to universities. Thus a Back-propagation network is applied to estimate the relative

important input variables. From the neural network, we can find the optimum values of the weights that minimize the error between the measured and the evaluated (output) performance parameters. Once the important input variables are retained and the rest are filtered out, a data mining technique – C4.5 decision tree, is modeled for generating the selection rules of admissions to universities as well as to build a knowledge base for storing up the decision-making rules. The decision rules generated by C4.5 are validated by measuring its performance. When it satisfactorily reaches the criterion set by the user, the Data Mining module is ready to use the new data for classification and prediction. The design of Classifier module is shown in Fig. 5.

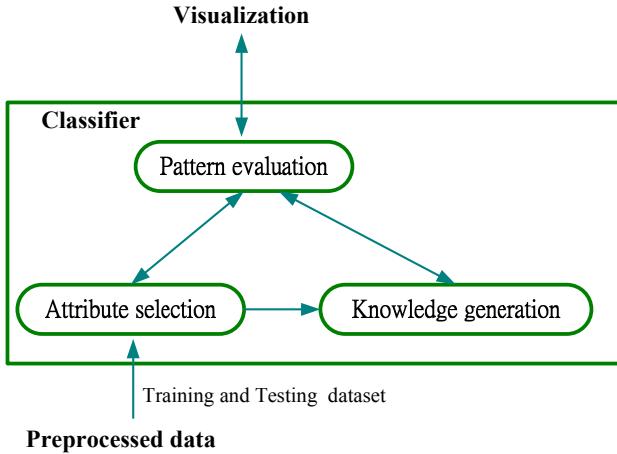


Fig. 5. The processing flow of the Classifier.

As an example shown in Fig. 6., the causal relationship can be extracted easily from the results of the decision tree model and represented in the form of classification IF-THEN rules. This is the process called knowledge generation. The leaf nodes are extracted according to the target output variables. Each If-THEN rule can be created on each path from the root node to a leaf node. Each attribute-value pair along a given path forms a conjunction in the rule antecedent ("IF" part). The leaf node holds the class prediction, forming the rule consequent ("THEN" part). The IF-THEN rules may be easier for humans to understand.

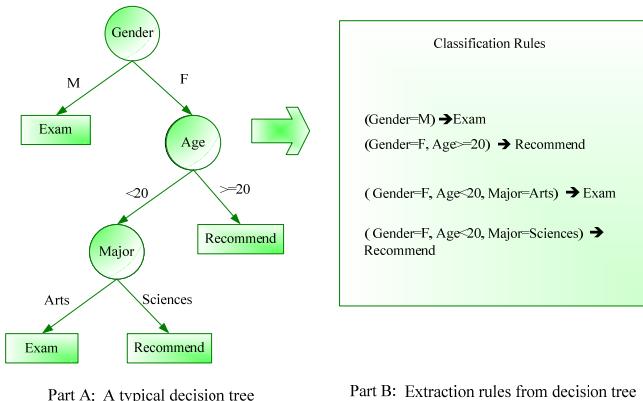


Fig. 6. Example of extracting rules from C4.5 decision tree.

A rule can be pruned by removing any condition in its antecedent that does not improve the estimated accuracy of the rule. For each class, rules within a class may then be ranked according to their estimated accuracy. Since it is possible that a given test sample will not satisfy any rule antecedent, a default rule assigning the majority class is typically added to the resulting rule set.

The Pattern evaluation component applies interestingness thresholds (Objective) set by users to filter out discovered patterns. At first, it employs thresholds and interacts with the Knowledge generation component to focus the search towards interesting patterns. The filtered rules (knowledge) are then encoded into knowledge base. The rules in knowledge base are then applied to new students to classify and predict their university admission.

Although objective measures (thresholds) help identify interesting patterns, they are insufficient unless combined with subjective measures that reflect the needs and interests of a particular user. For example, many patterns that are interesting by objective standards may represent just common knowledge. Subjective interestingness measures are based on user beliefs in the data. These measures find patterns interesting if they are unexpected (contradicting a user belief) or offer strategic information on which the user can act. Patterns that are expected can be interesting if they confirm a hypothesis that the user wished to validate, or resemble a user's hunch. So the Classifier provides a feedback for users to better discover interesting patterns and rules. After users applied the patterns and rules in Knowledge base, users can rank and score the patterns and rules by interestingness. So each rule with scores will be encoded together into knowledge base again, and they will be considered for later applying on new student data.

C. Visualization

The Visualization module communicates between the users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focusing the search, and performing exploratory data mining based on the intermediate data mining results. The Visualization module has the following features:

- Gain insight into an information space by mapping data onto graphical primitives.
- Search for patterns, trends, structure, irregularities, and relationships among data.
- Find interesting regions and suitable parameters for further quantitative analysis.

IV. EXPERIMENTAL EVALUATION

A prototype of RSAU system is implemented for experimental study. Real secondary students' data are used and the prototype program is written in Java and Weka. Both objective and subjective evaluations were conducted, by testing the performance of the hybrid classifier and feedback questionnaire survey done by human users respectively.

A. Classification Performance

Accuracy and error rate are the most common empirical measures. However, they do not distinguish between the numbers of correct labels of different classes for classifiers. Conversely, two measures that separately estimate a classifier's performance on different classes are *Precision* and *Recall*.

The *Precision* and *Recall* are two measures that separately estimate a classifier's performance on different classes. Technically, *Recall* is related to classification quality, and *Precision* is for the class-specific classification quality.

The *Precision* is the number of correctly classified samples for one class to the total number of samples classified to the class. In our case, for each predicted output Class_j, the *Precision* (*j*) is the ratio of the number of high graduate students (C_{jj}) were correctly predicted for enrolling university Class_j (e.g. University of Macau) to total number of high students classified for the university Class_j ($\sum_{i=1}^m C_{ij}$) in the student data set.

$$\text{Precision}(\text{for predicted Class}_j) = \frac{C_{jj}}{\sum_{i=1}^m C_{ij}} \quad (1)$$

The *Recall* is the number of correctly classified samples for one class to the total number of samples in the real output class. In our case, for each real output Class_i, the *Recall* (*i*) is the ratio of the number of high graduate students (C_{ii}) were correctly predicted for enrolling university Class_i (e.g. University of Macau) to total real number of high students in the university Class_i ($\sum_{j=1}^m C_{ij}$) in the student data set.

$$\text{Recall}(\text{for real Class}_i) = \frac{C_{ii}}{\sum_{j=1}^m C_{ij}} \quad (2)$$

Combining (1) and (2), we can derive *F1-Measure* which emphasizes the performance of a classifier on common and rare categories, respectively. Using *F1-Measure*, we can observe the effect of different kinds of data on a classification system.

$$\text{F1 - Measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (3)$$

A 10-fold cross-validation was used to carry out all the experiments, and averaged the results over 10 runs. In addition to the above criteria, we have tested the system using other criteria such as Classification costs, Learning performance, Explanation capability (c.f. [5]).

The following diagrams show the *F1-Measure* performances for the prediction of admission channels and classes of Universities, respectively in RSAU. From the experiment results, we can observe that our hybrid classifier model outperforms the other two. It achieved a high *F1-Measure* value ranging from 93% to 96% in both cases.

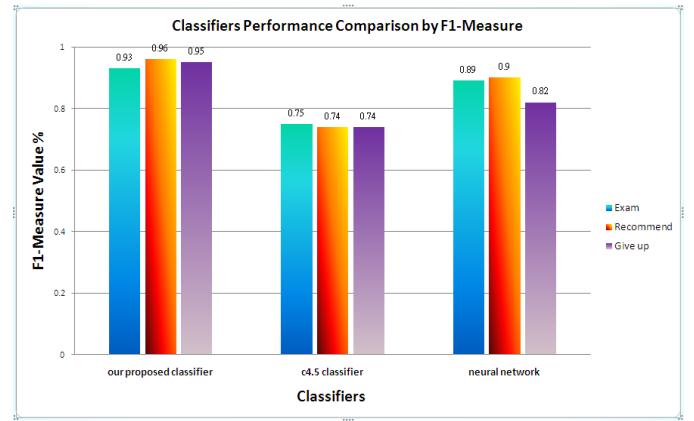


Fig. 7. The performance of classifiers in terms of *F1-Measure* of classes for Recommendation.

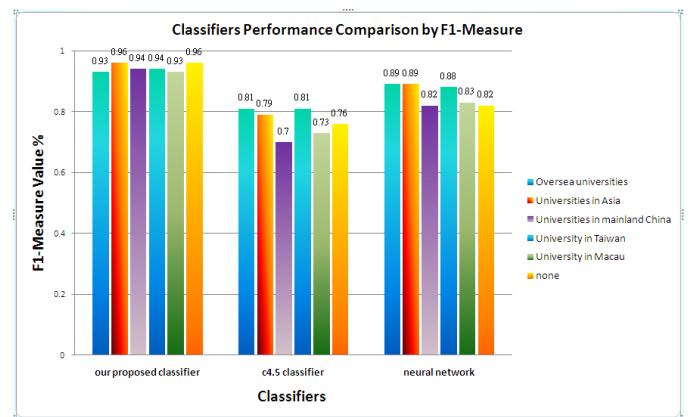


Fig. 8. The performance of classifiers in terms of *F1-measure* of classes for Admitted Universities.

B. Users' Feedback Evaluation

Usability is another measure to evaluate a recommender system by the actual users. In this project, a panel of users was invited for conducting a User Acceptance Testing over RSAU against the traditional manual method. The panel is comprised of 4 secondary school administration experts, 9 teachers and 9 students. The new system is more preferable in all aspects. Details can be found in [6].

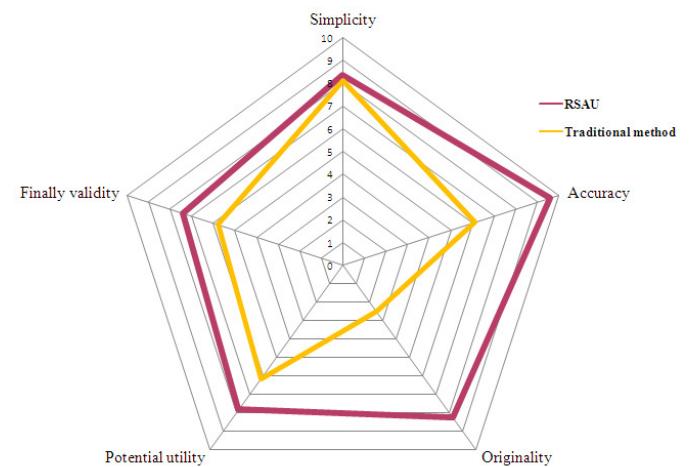


Fig. 9. Comparison on feedback evaluation of university admission studies by RSAU and traditional method.

V. DISCUSSION ON DECISION RULES

Other than predicting the chances of university admission, decision rules can be derived from the C4.5 decision tree in RSAU. A sample output fragment is shown in Fig. 10.

Rule 12 for Exam

```
Instances: 55      Confidences: 1.0
if Admitted_University = Universities in Taiwan
and Cga_ta > 75.800003
and Age <= 19
and Gra_ta <= 88.5
and Birth_place <= 1
then Exam
```

Fig. 10. A fragment of the importance decision rules for recommendation (with a high value of confidence).

When the decision rules are translated into a more comprehensive language, the rules do give some insights about the students and the university admissions. Some samples are derived in our experiments in the context of secondary schools in Macau. It is believed, however, RSAU would work equally well with data from students of secondary schools from other countries.

- Scores are still an important factor for qualifying students to universities, especially for Mathematics and English courses. When students' score of Mathematics and English are higher than 80%, almost all of them will be attended to universities. Whereas, when students' score of courses Mathematics and English are lower than 60%, they are likely to give up furthering their studies in universities.
- Female students, who are born in Macau are likely to enter local universities in Macau.
- The students who are major in Science class with youngest ages have higher probabilities of being recommended for direct admission to universities in mainland China.
- A large number of mature aged students have fewer chances to be recommended for direct admission to universities.
- In general, students who did not repeat in their last few years of secondary school study, have a good probability of admission to universities in mainland China or University in Macau.
- If the study status of a female student is not of any advancement and the average score is below 70%, the probability of giving up university study is higher than a male student.
- The female students from Guangdong have higher probabilities of giving up university study than those who are from Fujian, China.
- The female students who are major in Arts, and are born in China, have higher probabilities of recommendation to universities in China than those born in Macau.

VI. CONCLUSION

Education systems which do not have a standardized open exam for university admissions face the challenges of matching the right secondary school students with the right universities and the ways that they should enter. This implies some manual processes are needed and the decisions made are relied on human intuitions. Hence inaccuracies bound to happen.

In this project, we applied a hybrid data mining model to implement a recommender system prototype, called Recommender System of Admission to University (RSAU system). It analyzes various sources of secondary school students' data, in order to predict their chances of admissions to universities. It provides decision support about recommendations to university for secondary school administrators, teachers and senior secondary students.

The RSAU system had been evaluated by using real student data. The experiments showed that the hybrid decision tree and neural network approach improved accuracy in classification task. Although the real students' data used was from Macau, it is believed that the recommender is generic and applicable to educational systems in other countries such as those in [7].

ACKNOWLEDGMENT

The authors would like to thank Miss Lou Wan Chan, Ivy, for implementing the prototype of RSAU.

REFERENCES

- [1] J. C. Garcia and A. I. Zanfrillo, Data Mining Application to Decision-Making Processes in University Management, *INFOCOMP Journal of Computer Science*, volume 6, no.1 pp.57-65, 2007.
- [2] J. Luan, Data Mining Application in Higher Education, *SPSS Executive Report*, 2002.
- [3] J. Luan, Data Mining as Driven by Knowledge Management in Higher Education-Persistence Clustering And Prediction, *Keynote speech at the University of California-San Francisco's SPSS Public Roadshow*, 2001.
- [4] H. Jiawei and K. Micheline, *Data Mining: Concepts and Techniques*, Simon Fraser University, Morgan Kaufmann publishers, 2001.
- [5] S. Fong, Y. W. Si, R. P. Biuk-Aghai, "Applying a Hybrid Model of Neural Network and Decision Tree Classifier for Predicting University Admission", *The 7th International Conference on Information, Communications and Signal Processing (ICICS 2009)*, Submitted for publication.
- [6] W. C. Lou, "A Hybrid Model of Tree Classifier and Neural Network for University Admission Recommender System," Master of Science Thesis, University of Macau, Faculty of Science and Technology, 2008.
- [7] S. Alexander, M. Clark, K. Loose, "Case studies in admissions to and early performance in computer science degrees", *In ITiCSEWGR' 03: Working group reports from ITiCSE on Innovation and technology in computer science education*, ACM Press, pp.137-147