APPLY AND IMPROVE THE ASSOCIATION RULES IN PREDICTING NATIONAL HIGH SCHOOL ADMISSION SCORES TO NHA TRANG UNIVERSITY

Pham Thi Thu Thuy[1], Bui Xuan Huy[2], Truong Minh Hieu[3], Kim Hwa Soo[4]

[1] Nha Trang University,
[2] Khanh Hoa Education and Trainning Dept,
[3] To Van On High School - Khanh Hoa,
[4] AJOU University-Sout of Korea

Corresponding author:Pham Thi Thu Thuy; Email: thuthuy@ntu.edu.vn

Received: 15 Jan. 2024; Revised: 10 Feb. 2024; Accepted: 20 Mar. 2024

ABTRACT

The paper discusses how the traditional method of association rule mining based on user-defined minimum support and confidence can result in either too many or too few association rules. This could lead to valuable information being missed or redundant rules being generated, which is not practical and can be costly to implement. The paper proposes an improvement of DARIT algorithm to mine association rules without redundancy and applies it to data from previous years of National High School Exams to Nha Trang University. The goal is to use the results to build an admissions counseling system that can help students increase their chances of being admitted to universities based on their exam results.

Keywords*: Data mining, Association rule, Missing value, DARIT, consulting system.*

I. INTRODUCTION

The introduction Data mining is a technique based on the combination of computer science and mathematical theories, which is the process of automatically extracting valuable information hidden inside huge data sets in reality. Data mining has been integrated in many commercial data mining software and has many applications in many fields, especially managers, who use data mining as a support tool in the process of mining decision. Using this mined data, it is possible to predict national high school admission scores. With this data, the enrollment student numbers for less attactive university majors such as shipbuilding, fishing, fishery processing, aquaculture, etc., can be enhanced.

With the rapid development of computer networks, data is generated in extremely large quantities, today's analysts are having difficulty with the exploitation of big data, noisy data, data has constant and unstructured value. Meanwhile, users always want from the rudimentary data set in reality, they can find rules with high accuracy, easy to use, easy tounderstand, not too detailed and not duplicated in the rules. Rules must have high data coverage, contain valuable information, help managers make decisions that benefit the business. On the other hand, this process must have low hardware requirements and less time, and can be done at an acceptable cost.

Association rule mining is the process of finding all association rules in a database whose support (also known as popularity) is not less than user-defined minsup threshold and confidence is not less than user-defined minconf threshold.

The important issue is how the user should choose the threshold minsup and minconf to be able to generate the desired number of rules. This is an important issue because in reality users own finite resources (in terms of time and storage space). Depending on the choice of the minsup threshold, the current algorithms can be very slow and generate very large or very few association rules, the result set may miss valuable information or generate unexpected errors. redundant rules, have no use value in practice, take a lot of resources and costs to implement.

Agrawal [1] proposed the Apriori algorithm based on the idea of building association rules based on frequent sets satisfying minsup. The

frequent set is built on the principle: construct a candidate set of size k items that satisfy minsup threshold from frequent sets of k-1 items that satisfy minsup threshold. Apriori is simple, easy to understand and easy to implement algorithm. However, Apriori has disadvantages such as the database lacks some attributes, the algorithm has not yet proposed a solution to complete the data.

To overcome the above drawback of the Apriori algorithm, Grzegorz and colleagues [2] have proposed the DARIT algorithm to help complete the missing database based on association rules. Some other similar researchs [8-10] also proposed methods to process on missing data, but they did not mention the case of missing data. Number. So, it is necessary to improve their methods to exploit missing data in number form. In this paper, we choose to improve DARIT algorithm because it is close to our desired application.

Every year, in Vietnam, millions of candidates apply for entrance exams to hundreds of universities with many different fields and professions. For a candidate, in addition to choosing which school or major to take the exam to suit his or her interests, choosing which school to apply for admission is suitable for his or her own learning capacity to be able to take the exam. High enrollment also presents a big challenge.

A system to support decision making in which university to apply for admission to, which major to both suit their own interests and have the highest probability of admission for candidates is our main goal of this research.

II. MATERIAL AND METHODS

Semi-automatic admissions counseling support system [3] uses a combination of techniques in word processing, machine learning SVM and SMS processing in mobile communication systems. This counseling system is capable of receiving candidates' questions from the Web/email or via SMS. These questions are then automatically classified by SVM technique to be sent to the appropriate expert in each field. After having the answer from the expert, the system will give immediate feedback to the candidate. In addition, as soon as the candidate asks a question, the system will process and find the similarity of the current question with the previously answered questions, in order to suggest to the candidate more information. Testing on the data set collected from 447 questions in 8 areas that are often interested by many candidates shows that the system has an accuracy of 82.33%. This accuracy will also improve over time when the number of questions is large enough for the machine learning model, so this proposed solution will open a new direction in admissions counseling support.

Akos et al [4] proposed a system for counseling the college admissions. This study found that students and parents have high expectations for what college counseling should offer, but there is often a misalignment with what counselors can actually provide. Recommendations included improving counselor training, increasing counselor student ratios, and using technology to supplement counseling efforts.

Wu et al [5] developed an intelligent admission decision support system for universities. This study proposed an intelligent decision support system for universities to improve the admissions process. The system was designed to provide personalized recommendations for each applicant based on their academic history, extracurricular activities, and other factors.

Dell'Olio et al [6] proposed aA study on the effectiveness of early admission policie. This study examined the impact of early admission policies on student outcomes. The results suggested that early admission policies were effective in increasing the likelihood of enrollment for admitted students, but not necessarily in improving retention or graduation rates.

Lu et al [7] proposed a university admissions counseling through data analytics. They explored how data analytics can be used to

improve the university admissions counseling process. The authors found that data analytics can help identify patterns in applicant data, create personalized recommendations based on those patterns, and ultimately help improve the admissions process.

Overall, these studies suggest that there is room for improvement in the university admissions counseling system, and that technology and data analytics can play a role in improving the process for both counselors and students.

III. RESULTS AND DISCUSSION

1. Original DARIT algorithm

The DARIT algorithm [2] discovers association rules from incomplete transaction data, known as DARIT sequel (Discovering association rules in incomplete transactions). This approach allows incomplete transactions that can have any number of elements missing. The algorithm starts from describing the data structure, called mT-tree. DARIT algorithm is as follows:

1. mTd: mT-tree;
2. Apriori_Adapt(kmD, nkD, mTd)
3. for each transaction t *G* nkD
4. begin
5. Generate_Set_NZ(t, mTd);
6. mTd.Modify_probSup(t,t.NZ);
7. end
8. Generate_Rules(mTd);

At the beginning, the set of latent frequent itemsets is created (line 2) by calling the procedure ApriorLAdapt. In the process of determining likely frequent itemsets, pessimistic support and minimal pessimistic support are used. Minimum pessimistic support, also denoted minsup, is an additional parameter of the algorithm. This parameter defines the threshold that needs to be exceeded by the pessimistic support level of each item group in order to consider the item group as probable. It allows for an appropriate limitation of the number of sets to be considered during the execution of the Apriori-Adapt procedure, especially in the case of significantly incomplete data. In the next step of the DARIT algorithm for each incomplete transaction t, based on the tuples stored in the mT tree, a tuple t.NZ is generated; it includes sets that can appear in place of the null special element, indicating the element is missing from transactions (line 5). For each element of t.NZ, the probability of occurrence in the transaction under consideration is assigned. Based on the results obtained in this step, the values of probSup support of the frequent itemsets in the mT tree are modified (line 6). At the end of the algorithm, the GenerateRules procedure is called - it generates association rules from the mT tree using the estimated support values of the frequent sets.

Procedure Apriori_Adapt(Set of transaction kmD,nkD; mT-tree mTd)

1. Add_Frequent Atems(kmD, nkD, mTd)
2. p=2;
3. while(Generate_Candidates(p, mTd)> 0)
4. begin
5. for each transaction tGkmD
6. Calculate_Support(t, mTd);
7. for each transaction tGnkD
8. Calculate_Support_Incomplete(t, mTd);
9. mTd.Remove_NotFrequent(minSup, min_minSup);
10. p=p+1;
11. end

The Apriori-Adapt procedure starts from adding a likely frequent 1-itemset to the base of the mT tree (line 1). The Generate-Applications function generates candidate sets and returns their numbers. The candidate sets in a node n at level p are created by creating child nodes at level p +1, for each field in the tblElem table, except the last one. In the child node cn created for the jth field, the elements table includes all those elements from the parent node's tblElem table, stored in the fields of the index greater than j. The CalculSupport procedure increases the value of optimistic support and pessimistic support for those in the candidate pool supported by a complete transaction t. The CalculSupport-Incomplete procedure

(line 8) differs from the CalculateSupport procedure in that it increases the values of optimistic support for each candidate set. The RemoveNotFrequent method removes candidate sets that are unlikely to be frequent, and for the remaining candidate item sets, it sets the value of probSup support to the value of pessimistic support.

2. Modified DARIT algorithm

According to current admission regulations, the admission criteria of candidates to universities are affected by the following factors:

1. Enrollment quota of each school/major

2. Number of enrollments (trend of the major)

3. The quality of the test of the year that the candidate took the exam

Therefore, we propose a method to determine the university entrance scores of candidates based on the above factors as follows:

1. Enrollment quota given by schools

2. The number of registrations tends to increase or decrease according to the popularity of the major.

3. The quality of the author's exam using the test scores of candidates in Khanh Hoa province.

As a result from the data of previous years, we have the following information set (also known as rule set):

School year (nam)	School code (truongId)	Major code (nganhId)	Target (chitieu)	Benmark (diemchuan)	Pass rate (tyled)

The pass rate (Tyled) is calculated as follows:

1. Diemxt = max(Diemth1, Diemth2, Diemth3, Diemth4)

2. Tyled = (count(Diemxt>= diemchuan)/ tongts)*100

Problem: When a candidate wants to learn about a major of a certain university to participate in the entrance examination, the consultant needs to determine the following information:

School year	School code	Major code	Target	Benmark	Pass rate
N	X	Y	Z		

From the above table we see that the problem for the consultant is to identify the missing data (Benchmark and Pass Rate).

To determine the pass rate and benchmark in this algorithm, we improve the DARIT algorithm to find the missing data as above.

The university admissions for each subject includes scores of many blocks (subject combinations). When participating in the entrance examination, candidates will choose the combination of subjects with the highest score to register. This score is called the admission score.

For example, candidate has the following test results

Mathematics: 8.40 Literature: 6.00 English: 3.60 Physics: 7.25 Chemistry: 8.00 Biology: 5.75.

The candidate wants to apply for admission to Nha Trang University - Mechanical Engineering (Branch code 7510202)

This field of study has 4 combinations for admission: A00; A01; C01; D07".

Here are the score ranges for each combination:

A00: Math + Physics + Chemistry = 8.40 + 7.25 + 8 = 23.65

A01: Math + Physics + English = 8.40 + 7.25+ 3.60 = 18.65

C01: Literature + Math + Physics = 6 + 8.4 + 3.60 =18

D07: Math + Chemistry + English = 8.40 + 8 +3.6 = 20

In this case, candidates will choose block A00 to register for admission because this block has the highest score.

Procedure for determining admission scores (Diemxt) as follows:

1. Diemxt(int id, [Bind("table. properties")] DiemTohop diemTohop)

```
2. var ds = _context.DiemThis.ToList();
3. for each (var d in ds)
4. Diemxt xt = new Diemxt();
5. if (d != null)
6. if ((d.Mon1 == 0) | (d.Mon2 == 0) | (d.Mon3 == 0)) tohop = 0;
7. else tohop = (d.Mon1 + d.Mon2 + d.mon3);
8. xt.property = Max({ tohop1, tohop4, tohop3, tohop4 });
9. await _context.Diemxt.AddAsync(xt);
10. await _context.SaveChangesAsync();
```

The pass rate is the percentage of candidates with an entrance exam score > = the standard score out of the total number of eligible candidates (the admission score is greater than 0) into an examination industry. The procedure for calculating the pass rate is as follows:

```
1. Tinh_Ty_Le_Dat(int id, [Bind("Table. properties")] Luat luat)
2. var xt = _context.Luats.ToList();
3. var s = _context.Diemxts.ToList();
4. foreach (var d in xt)
5. if (d != null)
6. if (d.MaNganh == MaNganh)
7. var y = from t in s
8.    where Table.Property > 0
9.    select t;
10. var ts = y.Count();
11. var tl = from t in s
12. where Property >= d.DiemChuan
13.    select t;
14. var x = tl.Count();
15. d.Tyled = ((x / ts) * 100);
16. _context.Update(luat);
17. await _context.SaveChangesAsync();
```

The predicted pass rate includes the total pass rate of the latest year and the difference between the two most recent consecutive years. The algorithm for calculating the prediction success rate is as follows:

Tyled_du_doan()

```
1. var l2 = _context.Luat22s.ToList();
2. var l0 = _context.Luat20s.ToList();
3. var l1 = _context.Luat21s.ToList();
4. foreach (var d in l2)
5. if (d != null)
6. if (d.MaNganh == t.MaNganh and d.MaTruong == t.MaTruong)
7. var tl1 = from t in l0
8. where t.MaNganh == d.MaNganh and t.MaTruong = d.MaTruong
9. select t.Tyled;
10. var x1 = tl1.Max(x => x);
11. var tl2 = from t in l1
12. where t.MaNganh == d.MaNganh and t.MaTruong = d.MaTruong
13. select t.Tyled;
14. var x2 = tl2.Max(x => x);
15. if (x2 - x1 < 0) d.Tyled = x2 + (x2 – x1)/20;
16. Else d.Tyled = x2 + (x2 - x1) / 10;
17. if (luat.MaNganh != null)
18. context.Update(luat22);
19. await _context.SaveChangesAsync();
```

The predicted benchmark is the score converted from the predicted pass rate on the candidate's total score >0.

IV. CONCLUSION

By improving the DARIT algorithm to exploit the data of the National High School Exams 2020, 2021 and 2022 of Khanh Hoa province, with more than 40,000 exam data samples, the research has come up with a model to build the system. advising on admission to universities based on test scores of subjects included in the university admissions subject complex according to current regulations of the Ministry of Education and Training.

New point of TuVanTs algorithm (Enrollment consultant)

1. Provide a method to determine the entrance examination score on the basis of the candidate's transcript.

2. The method of calculating the pass rate is based on the benchmark and the admission score

3. Proposing a method to determine the predicted pass rate based on the pass rate of previous years

4. Predictive Benchmarking Method

Followings are a few results of our research. From the admission criteria (Target), the benchmark (Benmark), we calculate the

passing rate of the candidates (Pass rate). The matriculation rate is considered by discipline, in the framework of the graph, we illustrate the admission rate of 10 typical majors of Nha Trang University, the results of 3 years 2020, 2021 and 2022 are shown. presented in Fig 1, Fig2 and Fig3 respectively.

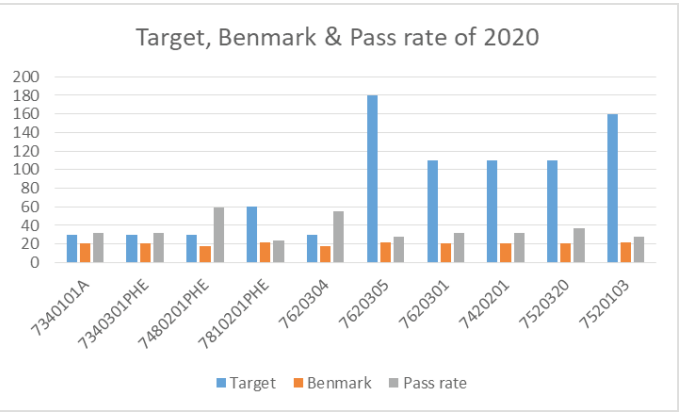

Figure 1 The given Target & Benmark and our Pass rate for the year 2020.

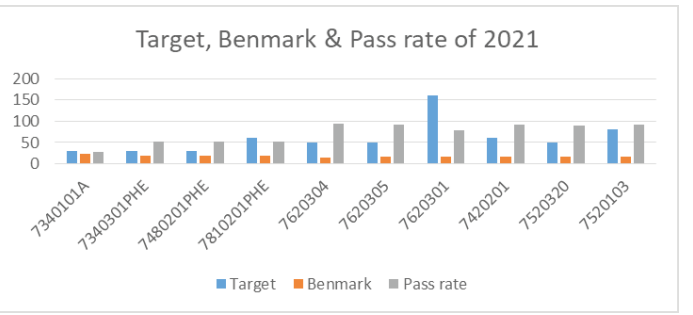

Figure 2 The given Target & Benmark and our Pass rate for the year 2021.

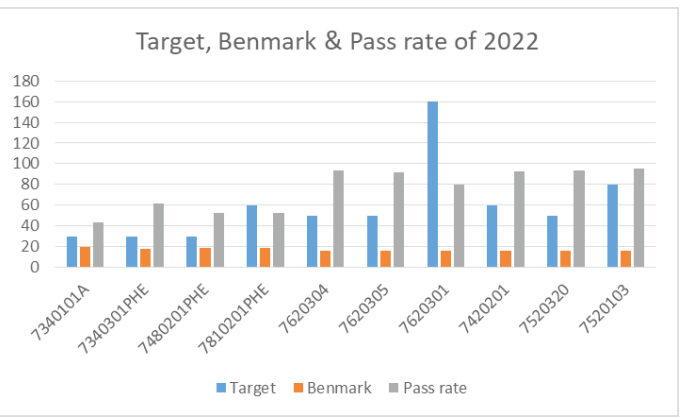

Figure 3 The given Target & Benmark and our Pass rate for the year 2022.

In addition, based on the Pass rate of the last 3 years (2020, 2021, 2022), we also predict the Pass rate of 2023 for the 10 groups of major mentioned above, the results are presented in Fig4.

Thus, with the above prediction results, the system can completely advise candidates to choose which major to apply for admission in order to achieve the highest passing rate.

In addition, the ability of candidates to

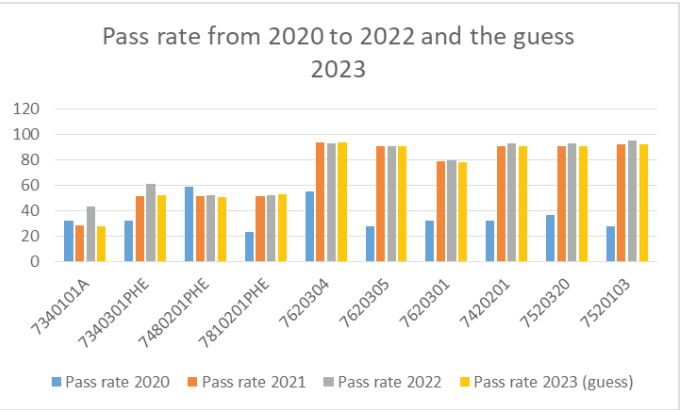

Figure 4 The Pass rate from 2020 to 2022 and our guess for 2023.

be admitted depends on how they apply for admission to the university. With the same test score, candidates applying to two different universities will have different chances of being admitted. With the above results, the consulting system can be completely applied to university admission counseling in practice.

The development direction of this research is to build a complete recommender system based on exploiting the database of exams large enough by the TuVanTs algorithm combined with the university's admissions benchmarks so far so that the system can the ability to advise all exam blocks, advise on which university to choose for the highest probability of matriculation. The recommender system in this study holds great potential for increasing the enrollment stundent in less attractive university majors, such as shipbuilding, fishing, fishery processing, aquaculture, etc.

REFETENCES

1. Agrawal, Rakesh, and Ramakrishnan Srikant, Fast algorithms for mining association rules", Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215 (1994).
2. Grzegorz Protaziuk, Henryk Rybinski, Discovering Association Rules in Incomplete Transactional Databases, Transactions on Rough Sets VI, Volume 4374 (2007).
3. Nguyen Thai Nghe, Truong Quoc Dinh, University Admissions Support System, Scientific Journal of Can Tho University, IT Topics 2015, pp. 152-159 (2015).
4. Akos, P., Schulenberg, S. E., & Lodes, S, An exploration of college admissions counseling expectations and reality, Journal of College Counseling, 19(1), 50-63. Link: https://doi.org/10.1002/jocc.12019 (2016)
5. Wu, K., & Tang, L. (2018). Developing intelligent admission decision support system for universities. IEEE Access, 6, 5169-5179. Link: https://doi.org/10.1109/ACCESS.2018.2793843
6. Dell'Olio, L., & Szydlowski, M. (2018). A study on the effectiveness of early admission policies. Journal of College Admission, (238), 22-29. Link: https://search.proquest.com/docview/2158766721
7. Lu, M., & Zheng, J. (2019). Improving university admissions counseling through data analytics. International Journal of Emerging Technologies in Learning (iJET), 14(2), 71-80. Link: https://doi.org/10.3991/ijet.v14i02.9344
8. K. Rameshkumar, A novel algorithm for association rule mining from data with incomplete and missing values, ICTACT Journal on Soft Computing, Vol 01, Issue 04 (2011)
9. Tzung-Pei Hong and Chih-Wei Wu, Mining rules from an incomplete dataset with a high missing rate, Journal: Expert Systems with Applications, Volume 38, Number 4, Page 3931 (2011).
10. Zhigang Sun, Mengmeng Gao, Aiping Jiang, Min Zhang, Yajie Gao, Guotao Wang, Incomplete data processing method based on the measurement of missing rate and abnormal degree: Take the loose particle localization data set as an example, https://doi.org/10.1016/j.eswa.2022.119411 (2023).