Data Mining of Students' Course Selection Based on Currency Rules and Decision Tree

Yu Liang, Xuliang Duan, Yuanjun Ding, Xifeng Kou, Jingcheng Huang
Sichuan Agricultual University,Yaan Sichuan 625000,CHINA
Key Laboratory of Agricultural Information Engineering of Sichuan Province,0086-6250
86-15984661614; 86-15008305394; 86-17759720996; 86-17321919306; 86-17321910847
1371263005@qq.com; duanxuliang@sicau.edu.cn; 1351265426@qq.com;
984425976@qq.com; 910771551@qq.com

ABSTRACT
The currency of data can ensure that data is not obsolete and outdated. As one of the important bases for evaluating data quality, it plays an important role in the availability of data. Data currency rules can effectively discriminate the temporal relationship between data sets. The decision tree can availably classify and predict the data, and can test the attribute values very well. In this paper, the currency rules are combined with the C4.5 algorithm in the decision tree, and the improved algorithm is applied to the college elective data in recent years. Through experiments, the algorithm used in this paper can extract the statute rules from the student elective database. According to the currency rules, the college teaching plan can be planned in advance and the curriculum resources can be allocated reasonably.

CCS Concepts
• **Information systems → Information systems applications → Data mining → Association rules**

Keywords
currency rules; decision tree; course selection information; data mining

1. INTRODUCTION
Currency refers to the difference in the nature of the same thing at different times. The currency of data means to ensure that data keeps pace with the times and is not obsolete. As the scale of universities continues to expand, how to improve the rational allocation and use of teaching resources through teaching management has become a close concern of universities. The elective system means that college students are allowed to have certain freedom of choice for the courses offered by the school, including selecting courses, instructors and class time, and choosing the appropriate amount of learning and learning process[1]. Each year, students' elective behavior will generate a huge database. After the course selection, there will be a large amount of elective information redundancy, which is not fully utilized. By analyzing the time-sensitive relationship between the elective data in the database, extracting valuable information, the main factors affecting students' electives can be found. There may be links in these data, for example, students of the same profession are more likely to choose relevant professional electives. In addition, information is not long-lasting, and some information will lose its effectiveness after a period of time. Such as the graduated students, their course information will become invalid. In order to solve the above problems, this paper uses the currency rules and the C4.5 algorithm in the decision tree to filter the course selection information, and effectively clean, filter and analyze the course selection information. The useful information is fed back to the instructor so that the instructor can arrange the course resources in advance and reasonably. Under the environment of rapid development of networked education, making full use of data mining technology has important practical significance for the reform of teaching management in universities.

ICBDC 2019, May 10–12, 2019, Guangzhou, China

ACM ISBN 978-1-4503-6278-8/19/05…$15.00
http://doi.org/10.1145/3335484.3335541

2. TECHNICAL FOUNDATION
This paper mainly judges the data state type currency rules for students' course selection data. In order to better describe the currency of data, we first introduce the definition of the concept related to the definition of the problem. Through the currency rule and the decision tree C4.5 algorithm, the class information is classified, the data is filled, cleaned and modified, and the effective screening is selected. information. The specific architecture of the system is shown in Figure 1.

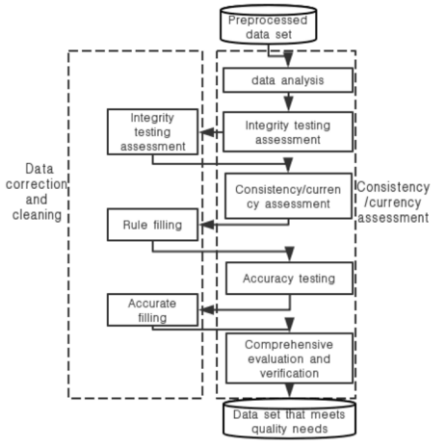

Figure 1. System architecture design.

2.1 Currency Rules

In a broad sense, the currency rule of data refers to the quality changes that result from the influence of time. The study of time-dependent correlation properties in literature [2-3] determined that currency was mainly affected by the three dimensions of data freshness, data decay, timeliness and availability. According to domestic and foreign research, the current research on the currency rule of big data can be summarized as the following three types of methods: currency judgment based on time stamp; rule-based currency judgment and relative data currency judgment.

Currency judgment based on time stamp refers to judging whether the record is outdated by querying fresh data records within a certain transaction time. Rule-based currency judgment refers to the use of denial constraints and copy functions between different data sources [4] in the absence of timestamps in the data. Deducing the currency of data describing the same entity record in the same data set. Literature [5] proposes the latest value query and user related currency. Relative data currency judgment refers to the relative currency order determination of records and attribute values describing the same entity[6]. Literature [7] is mainly the research work on relative data availability, which lays the foundation for the study of relative data currency. After analyzing the characteristics of the data collection of students' course selection, this paper uses the rule-based currency judgment to evaluate the data, which is convenient for later data cleaning, correction, screening and analysis.

The definition of rule-based currency is as follows: Firstly, according to the characteristics of attributes, we define the currency rule as: some attributes show increasing, decreasing or obeying a specific state transition sequence in the time order of tuples, which we call the attribute satisfying the timeliness rule. Secondly, the database schema relational database schema containing m attributes is represented as R=(EID, A1, A2, A3,..., Am), where EID is the entity identifier, and different tuples with the same EID correspond to the same entity (entity), EID can be generated by entity recognition technology [8,9]; Ai is used to represent the i-th attribute, and dom(Ai) is the value range corresponding to Ai.

The currency rule is defined as follows: On a relationship R, if it is determined that the state currency rule between different tuples belonging to the same entity, then the tuple attribute and value type currency rule must be determined first[10]. On the relationship R, for the tuple (t_1, t_2) of the same data entity, if some attributes are sequentially incremented, decremented, or subject to a specific state transition sequence on the chronological axis of the tuple, the function dependencies are not violated. Under the premise of conditional dependencies, the following relationships are satisfied:

$$\forall t_1, t_2 \in R, (t_1[EID] = t_2[EID] \wedge t_1[Ai] \prec t_2[Ai]) \rightarrow t_1 \prec t_2$$

A certain numerical attribute Ai satisfies: the state time validity judges the process state according to its attribute $v_1 \rightarrow v_2$ (the state diagram expressing the attribute currency rule is a directed graph G(V, E), where the vertex V is the attribute value (state) a finite set, the directed edge set E represents the direction of the chronological state transition.) In its state diagram, if the tuples t_1, t_2 have the attribute values v_1, v_2 and v_1, v_2 are the state nodes of the graph G, if t_2 is newer than t_1, the state currency rule is expressed as:

$$\forall t_1, t_2 \in R, t_1[EID] = t_2[EID], v_1 = t_1[Ai], v_2 = t_2[Ai], v_1 \in G, v_2 \in G$$
$$t_1 \prec t_2 \rightarrow (v_1 \rightarrow v_2 \wedge \neg v_2 \rightarrow v_1)$$

That is: in the state diagram G, v_1 to v_2 are reachable and v_2 to v_1 are unreachable.

Similarly, on a consistent data set:

$$\forall t_1, t_2 \in R, t_1[EID] = t_2[EID], v_1 = t_1[Ai], v_2 = t_2[Ai], v_1 \in G, v_2 \in G$$
$$(v_1 \rightarrow v_2 \wedge \neg v_2 \rightarrow v_1) \rightarrow t_1 \prec t_2$$

The student's elective status in different semester can be used as a directed graph as shown in Figure 2:

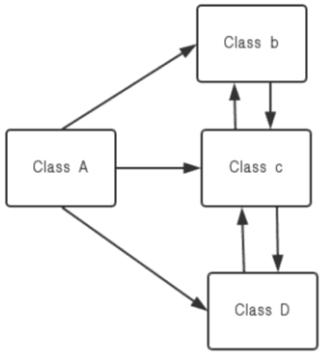

Figure 2. shows a directed graph of attribute state changes.

If there is no loop between the two states, the two states have a certain temporal relationship; if there is a loop, the temporal relationship is uncertain [11].

Thirdly, combined with the characteristics of college students' course selection database, this paper defines the currency rules as follows: On the same dataset, for a state currency rule with the same kind of attributes, it satisfies the number of statute rules and the rules of satisfaction and violation of currency. The ratio of totals is called support.

$$s_c = \frac{Cnt(A -> B)}{\mathrm{Cnt(A -> B)} + \mathrm{Cnt(B -> A)}}$$

Where S_C denotes the support degree of rule C, Cnt(A->B) denotes the number of arrivals of A to B, and Cnt(B->A) denotes the number of arrivals of A to B [10]. In the specific students' course selection study, according to the directed graph of the student's elective state, the number of students' electives is represented by the weight of the corresponding side in the directed graph. As shown in the figure 3, the weight from B to C is 90, and the weight from C to B is 10, then the support degree of rule B to C is S(B→C) =|90|/ (|90|+|10|) =90%, Similarly, S(B→D) =46%. According to the algorithm, for a specific state currency rule, the stronger the currency rule, the greater the support degree.

2.2 Decision Tree

Decision tree algorithm essentially uses a series of rules to classify data. The generation of tree refers to the process of recursively generating decision tree from the root node in the process of continuous traversal and continuous selection of the optimal scheme. In the process of traversing the decision tree, the change of each node may lead to dramatic changes in the results, which is to say, the change of a decision node may lead to the change of leaves [11]. This paper mainly uses the most commonly

used C4.5 algorithm of decision tree classification algorithm to study and analyze the data of students' course selection.

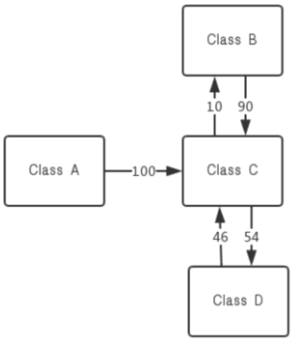

Figure 3. Calculation of support for currency rules in directed graphs.

C4.5 algorithm is a more effective and accurate algorithm based on ID3 algorithm. By choosing split attributes based on information gain rate, we overcome the shortcomings of ID3 algorithm in choosing attributes with multiple attribute values as split attributes. The goal of C4.5 is to find a mapping relationship from attribute values to categories by learning, and this mapping can be used to classify new entities with unknown categories. The classification rules generated by C4.5 algorithm have the advantages of easy understanding and high accuracy. The flow chart of decision tree C4.5 algorithm is shown in Figure 4：

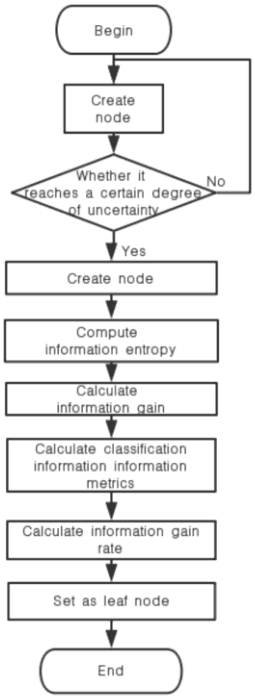

Figure 4. Decision Tree C4.5 Algorithm Flow.

3. THE SPECIFIC IMPLEMENTATION

3.1 Thinking Process

The degree of support in the currency rule can be used to determine the strength of the temporal relationship between two attributes. In the same way, the use of support degree can be used as the basis for the temporal strength between the courses of the students during the course selection. Using the nodes of the decision tree, students can quickly classify the course data according to the semester. Each layer can represent a semester, and each node represents a course to construct a decision tree. In the process of traversing the decision tree, in the tree structure of the decision tree, each node represents a test on an attribute, each branch represents the characteristics of a test output, and the weight of each node is calculated, and the support degree is used as the weight. To evaluate the relationship between the two nodes. In each traversal process, the formula is used to calculate the support between each two layers. Finally, the support degree in each course selection process can be output. Using decision tree classification and traversal, and calculating the support between attributes, you can get the support strength between each course selection process.

In the decision tree of Figure 5, we can get [1, 2, 4, 7, 8], [1, 2, 5, 7, 8], [1, 2, 6, 7, 8], [1, 2, 4 , 7, 9] and so on a total of 12 cases, for each case, you can find the support degree of two adjacent courses A-> B, specifically use the support degree formula Cnt (A-> B) / (Cnt (A ->B)+ Cnt(B->A)), the condition that the support degree of A->B is less than 1 if and only if B->A exists, and the more the number of B->A exists, the smaller the support degree of A->B, the currency indicated the weaker. Finally, the data set is put into the C4.5 algorithm, and the course selection order related to the time-dependent relationship can be determined. The decision tree structure diagram is as shown in Figure 5 below.

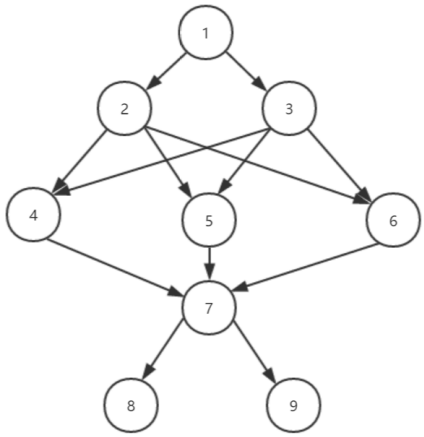

Figure 5. Decision tree structure.

3.2 Data Processing

The data processing in this paper mainly includes data cleaning and data preprocessing of the student elective data set.

Data cleaning refers to splitting each student's selected course according to different semester, and using the new set obtained by splitting as the node set of the decision tree. The data preprocessing is to improve the readability of the algorithm, and the key selection is used to not number all the selected courses in the whole school. The relative sequence of data in the collection represents the semester order. In the collection {A, B, C, D...}, A means that the student chooses the A course in the first semester,

B means that the student chooses the B course in the second semester, and so on.

The time relationship of the course selection is determined by the position of the position in the collection. According to the above data processing, the student course information map shown in Table 1 can be constructed:

Table 1. Students' course selection information diagram

Student name	Term1	Term2	Term3	Term4	...
Name1	A	B	C	E	...
Name1	A	B	D	E	...
Name2	F	B	C	G	...

3.3 Algorithm Design

The algorithm design is mainly for each student to select the course of the course in different semester, use the decision tree to traverse all the courses, and calculate the support degree of each of the two courses for each student. The algorithm flow assumes that one tree represents a student's elective situation. Each level represents one semester, and the number of nodes in each layer represents the number of courses selected in the current semester.

Algorithm 1: Course Division

```
for each entity list in cur_list do
 for each value course in all_record[num] do
  list1←list[:];
   if course not in course_id.keys() then
      course_id[course]←course_num;
      course_num←course_num+1;
      list1.append(course);
      list2.append(list1);
  endfor
endfor
```

For each entity, first iterate through all the cases and assign all traversed lessons to a unique ID. Replacing the corresponding course with an ID number can make the algorithm run more efficiently. After traversing, all the permutations and combinations are output, and the ID is assigned to these courses, so that the support degree can be calculated later.

Algorithm 2. Define timing relationships

```
for each entity list in cur_list do
  data_set.append(list);
   for each_id←0to len(list)-1 do
    tmp_s←str(list [each_id])+"->"+str(list[each_id+1]);
       if tmp_s in support.keys() then
          support[tmp_s]←support[tmp_s]+1;
       else
          support[tmp_s]←1;
      endif
   endfor
endfor
```

Before calculating the support degree for the two courses, it is necessary to define and count the relationship of the relationship between each two courses.

Algorithm 3. Calculate the support of two adjacent courses

```
for each value S in support do
  a←int(S.split('->')[0]);
  b←int(S.split('->')[1]);
 s←str(b)+"->"+str(a);
  if s in support.keys() then
    rate←support[S]/(support[S]+support[s]);
    else
      rate←1;
   endif
  support_rate[S]←rate;
endfor
```

For calculating the degree of support between each adjacent course, first define a to indicate the number of violations of the currency rule in the entity, and b to indicate the number of currency rules in the test. If there is a temporal relationship in the entity S that does not satisfy the currency rule, the support degree is used to calculate S. If the temporal relationship is satisfied in the entity, the corresponding support degree can be expressed as 1.

Algorithm 4 calculates the average support degree for each entity

```
for each entity list in data_set do
 sum←0;
 length←len(list);
  if length==1 then
   continue;
  endif
   for each←0to len(list)-1 do
     tmp_s1←str(list[each])+"->"+str(list[each+1]);
     sum←sum+support_rate[tmp_s1];
   endfor
 ave←sum/(length-1);
  if ave>=0.85 then
list.append(0);
  else
     list.append(1);
  endif
endfor
```

For each entity, need to calculate the average of the support degree for all the attributes it contains. In order to be more persuasive, the average support degree is obtained for each course selection sequence, and the sum of the support degrees of the two adjacent courses is divided by the sequence length minus one. Identified the elective sequence with an average support of less than 0.85 as a sign of weak currency, and added a mark at the end of the sequence, with 0 indicating weak currency and 1 indicating strong currency.

4. TESTING AND APPLICATION

Experimental configuration: The experimental hardware environment is Intel(R) i5 6300 2.30GHz CPU, 8G memory, operating system Window10, program, this experiment is implemented in Python language.

The experimental data selected the students' course selection information of the 2015-2018 school year of a university as the application data set. The data set covers about 35,000 sample data, including the basic attribute set {college, profession, class, name, semester, choose course}.

4.1 Experimental Results

Due to the large number of data sets selected for the experiment, only some experimental results are intercepted in this paper. Figure 6 shows the experimental results from line 5 to line 60:

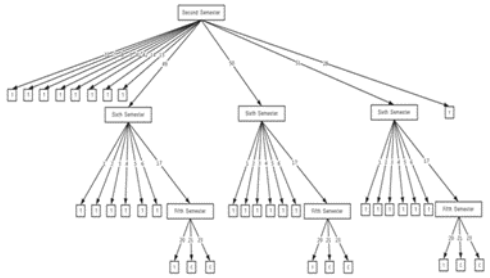

Figure 6. Screenshot of the experimental results decision tree.

In the experiment, each entity uses a decision tree to display intuitively. Starting from the second semester of the students' elective courses, the weight between each node represents the course number, and each node represents the support degree between each node and the adjacent courses. In this paper, we define the support degree greater than 0.85 to indicate that there is a strong currency between courses, expressed in 1. When the support degree is less than 0.85, it means that the currency between courses is weak, and the currency is not obvious, expressed by 0. For example, in Figure 7, the currency of curriculum sequence 49→6, 50→17→20 is strong, while the currency of curriculum sequence 49→17→23, 51→17→20 is weak.

Visualizing the dataset support obtained from the student's course selection data, can get the following results:

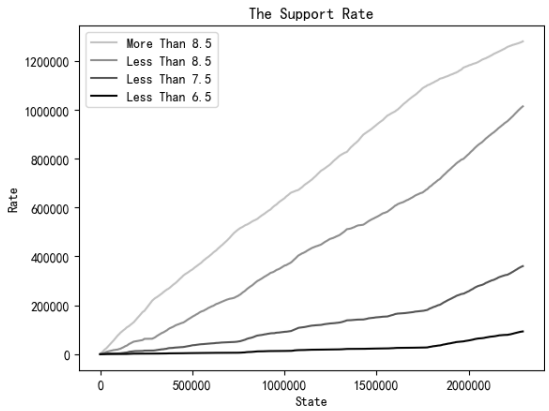

Figure 7. Line chart of data support degree.

It can be seen from Fig. 7and Fig. 8 that as the number of entities continues to increase, the gap between the support degree levels of entities is becoming more and more obvious, and the proportion of support exceeding 8.5 is the highest, and the entity with support less than 6.5 is extremely Less. In data support, the proportion of support greater than 8.5 also exceeds 50%. The proportion of support between 7.5 and 8.5 is also close to 30%. It can be intuitively felt from the figure that from the intuitive feeling in the picture, there is an obvious time sequence relationship in the process of students' choose courses, that is, there is also a time sequence in the choice between courses and courses. The larger the data set, the more obvious the rule is. It can be concluded that for most students of the same major, there are similar courses selection paths in the course of course selection, which belong to the progressive relationship of layers.

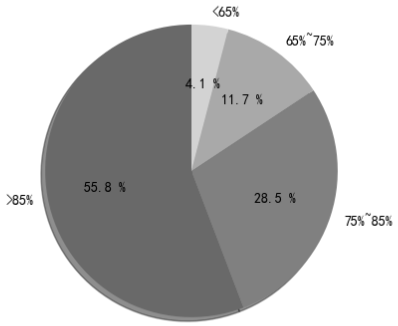

Figure 8. Pie chart of data support degree.

4.2 Experimental Analysis

In the course of course selection, one is to choose courses related to one's major, and the other is to follow one's own interest. The choice between courses has a certain degree of relevance, such as the complementary relationship between the two courses, or the inheritance relationship. If a student has two courses in any two adjacent semesters in all semester, it shows strong support and also shows strong currency. This shows that the student's course selection is correct and helpful. of. If the student's average support for all semester is less than a certain value, then the student's elective course is not only good for his or her major, but increases the academic burden.

After the experiment, we can make full use of these data sets with support degree greater than 8.5 to extract the temporal relationship of students' choosing courses to predict the trend of next students' choosing courses. According to the data set of course selection in the previous period, the Educational Administration Department of colleges and universities can determine the main course range of the corresponding majors in the next semester, so as to better arrange the course resources in advance. At the same time, it can also analyze the average support of all students in Colleges and universities, and determine the proportion of students in the normal development, as well as the proportion of students in the course of random selection and re-study. Through the analysis of currency, we can give some suggestions to the construction of school style of study and the reform of elective course system, so as to better promote the healthy development of students.

5. CONCLUSION

The currency algorithm can be used to judge whether the data is valid or not. Through certain data cleaning and sorting, and using algorithms to study the currency rules, a lot of useful information can be extracted, which can not only analyze students' course selection behavior, but also provide a reasonable basis for the next reasonable arrangement of educational and teaching resources. The experimental results show that the decision tree and currency rules can be applied to the course selection management in colleges and universities, and have a high degree of applicability. The rule extraction of students' course selection data can reasonably improve the quality of the use efficiency of the

school's network course selection system and give full play to the functions and effects of the system.

6. REFERENCES

[1] Feng Kepeng. *Research on Personalized College Course Selection Recommendation Algorithm Based on Data Mining[J]*. Digital Technology and Application,2012,(8):63-64.

[2] Sidi F, Shariat PPH, Affendey LS, Jabar MA, Ibrahim H, Mustapha A. *Data quality: A survey of data quality dimensions.* In: Proc. of the 2012 Int'l Conf. on Information Retrieval & Knowledge Management. IEEE, 2012. 300-304. DOI= http://doi.acm.org/10.1109/InfRKM.2012. 6204995.

[3] Fan W, Geerts F, Wijsen J. *Determining the currency of data.* ACM Trans. on Database Systems, 2012,37(4):25-41. DOI= http://doi.acm.org/10.1145/2389241/2389244.

[4] Dong X L, Berti-Equille L, Srivastava D. *Truth discovery and copying detection in a dynamic world[J]*. Proceedings of the VLDB Endowment, 2009, 2(1): 562-573.

[5] Jian-Zhong L I M H L I, Hong GAO. *Evaluation of Data Currency[J]*. Chinese Journal of Computers, 2012, 11: 013.

[6] Gao Yitong. *Research on Key Technologies of Big Data Timeliness [D]*. Heilongjiang: Harbin Institute of Technology, 2016.

[7] Nguyen H T H, Cao J. *Trustworthy answers for top-k queries on uncertain Big Data in decision making[J]*. Information Sciences, 2015, 318: 73-90.

[8] Zhang H, Diao Y, Immerman N. *Recognizing patterns in streams with imprecise timestamps.* Elsevier Science Ltd., 2013. DOI= http://doi.acm.org/10.14778/1920841.1920875.

[9] Fan W, Geerts F, Tang N, Yu W. *Conflict resolution with data currency and consistency.* Journal of Data and Information Quality(JDIQ), 2014,5(1-2):6. DOI= http://doi.acm.org/10.1145/2631923.

[10] Duan Xuliang, Guo Bing, Shen Yan, Shen Yuncheng, Dong Xiangqian, Zhang Hong. *Data Restoration Method Based on Time-Related Rules.* Journal of Software, 2019, 30(3): 589−603. http:// www.jos.org.cn/ 1000-9825/5688.htm

[11] Guo Qiaochi, Yang Hong. *Application of C4.5 Classification Decision Tree in College Course Management[J]*.Computer Knowledge and Technology,2018,14(2):249-251.