Study and Application of Apriori Algorithm in Students' Behavior of Taking Courses

Jing LIN

School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan,China

whpu_linjing@163.com

Keywords: data mining; association rule; Apriori algorithm; take courses

Abstract. Association rule mining is the core technology of data mining. The Apriori algorithm is introduced and applied to the process of students taking courses and teaching themselves by a network teaching system. On the basis of large amounts of data, frequent itemsets will be found by the Apriori algorithm. The frequent itemsets describe the relation between students' characteristics and their behavior of taking courses. According to these relations, when a new student first takes courses, the teaching system can intelligently predict his/her tendency of taking courses and recommend appropriate courses to him/her.

Introduction

With the development of database technology, data mining technology emerged. Data mining is the core technology of knowledge discovery in database. Data mining extract some knowledge from the database. This knowledge is implicit, unknown and potential information. Association rule mining is an important branch of data mining. The purpose of association rule mining is finding out useful relations between different items in large transaction database. These potential relevance and behavior patterns can help people make many business decisions, such as customers shopping analysis, catalog design, products advertisement mailing analysis, etc.

In recent years, network teaching system is widely used in university education. Students can study independently through the network teaching system. A network teaching system often provides many curriculum resources. Students can take courses that they are interested in and download course documents, or carry out other learning activities. The first time, most students are often blind when taking courses in the network teaching system. Association rule mining is applied to this issue to discovery some relations between students' learning behavior and characteristics. It makes the system recommend courses to students intelligently. Students can better study independently.

Association rule and Apriori algorithm

Association rule. For relational data or other information, association rules mining finds out frequent patterns among item sets or object sets, analyze relationships in data, judge what things will happen together.

I={i1,i2,…,in}. I is an itemset and i_n is an item. D is a set of transactions. Each transaction in D is a set of items, denoted by T($T \subseteq I$).

X is an itemset. T contains X($X \subseteq T$).An association rule is a formula like $X \Rightarrow Y$($X \subseteq I, Y \subseteq I, X \cap Y = \emptyset$).

In a transaction set D, the Support of rule $X \Rightarrow Y$ is the ratio of the number of transactions containing both X and Y and the number of all transactions, denoted by Support($X \Rightarrow Y$).

Support($X \Rightarrow Y$)=|{T:$X \cup Y \subseteq T, T \in D$}|/|D|

In a transaction set D, the Confidence of rule $X \Rightarrow Y$ is the ratio of the number of transactions containing both X and Y and the number of transactions containing X, denoted by Confidence($X \Rightarrow Y$).

Confidence($X \Rightarrow Y$)=|{T:$X \cup Y \subseteq T, T \in D$}|/|{T:$X \subseteq T, T \in D$}|

© 2016. The authors - Published by Atlantis Press

The Support represents a frequency of pattern in the rule. If s% transactions contain $X \cup Y$ in D, the Support of rule $X \Rightarrow Y$ is s. The Confidence represents credibility of a rule. If c% of transactions containing X also contain Y in D, the Confidence of rule $X \Rightarrow Y$ is c. The Support and Confidence of these association rules must be greater than the predefined minimum Support and minimum Confidence. This problem can be decomposed into two sub-problems.

(1)Get frequent itemsets;

(2)Get association rules from frequent itemsets;

The first step decides the performance of association rules mining. Most of mining algorithms focus on how to get frequent itemsets. Apriori algorithm is the first algorithm to solve this problem. In this paper, we discuss the application of Apriori algorithm.

Apriori algorithm. The core idea of the Apriori algorithm is to compute the frequent itemsets in the database. The frequent itemsets will be chosen from candidate itemsets step by step. Firstly, we scan the database to generate the first candidate itemsets. We count the appearance of each candidate data item in the database. Then the first frequent itemsets L_1 generate. In L_1, the Support of each data item must be greater than the minimum Support. Similarly, the second frequent itemsets L_2 will generate from the first frequent itemsets L_1. The calculation will stop until the n-th frequent itemsets L_n generate and the the (n+1)-th frequent itemsets L_{n+1} are unable to generate. The process of generating L_{k+1} from L_k can be divided into two steps.

(1)Connection. L_k connect with itself to generate the K+1 candidate itemsets, denoted by C_{k+1}. In an itemset, items are sorted by the alphabet order. L_1 and L_2 are both itemset in the k-th frequent itemsets L_k. $L_i[j]$ represents the j-th item of L_i. If the first k-1 items of L_1 are the same as the ones of L_2, the L_1 and L_2 can be connected to generate the candidate itemsets $C_{k+1}(L_1[1],L_1[2],\ldots L_1[k],L_2[k])$. The procedure is expressed by formal language as follows:

If $(L_1[1]=L_2[1]) \wedge (L_1[2]=L_2[2]) \wedge \ldots \wedge (L_1[k-1]=L_2[k-1]) \wedge (L_1[k]<L_2[k])$

Then $C_{k+1}=L_1 \bowtie L_2$

(2)Pruning. The candidate itemsets C_{k+1} contains the (k+1)-th frequent itemsets. In any sub-itemset C of C_{k+1}, all the k-th sub-itemsets are calculated. If there is a k-th sub-itemset which is not a frequent itemset, C will be deleted from C_{k+1}.Then we compute the Support of each itemset in C_{k+1}. Some Supports of itemsets are greater than the minimum Support. Those itemsets compose the (k+1)-th frequent itemsets L_{k+1}.

The following is the Apriori algorithm programming.

Input: D(a set of transactions) Minsup(the minimum Support)

Output: all the frequent itemsets

```
Lt={frequent1-itemsets};
  for(k=2;Lk-1≠∅;k++)
  {
      Ck=Apriori_gen(Lk-1);
      for each transaction t∈D
      {   //scan the database and calculate the Support
          Ct=subset(Ck,t);      //get the subsets of transaction t (these subsets are candidate sets.)
          for each candidate C∈Ct do
              c．count++；
       }
      Lk={c∈Ck|c.count≥Minsup}
   }
  return =∪Lk;
```

The function of procedure Apriori_gen is to generate candidate set and prune. Pseudo code is as follows:

```
for each itemset l1∈Lk-1
for each itemset 12∈Lk-1
```

if ($L_1[1]=L_2[1]$)∧($L_1[2]=L_2[2]$) ∧…∧ ($L_1[k-2]=L_2[k-2]$) ∧ ($L_1[k-1]<L_2[k-1]$)
then
{
$c=l_1 \bowtie 1_2$;
if has-infrequent-subset(C,L_{k-1}) then
delete c;
else
add c to C_k;
}
return C_k;

Association rule mining in taking courses

Now more and more universities develop a network teaching system to assist teaching. Many course documents are published via the system. Students can learn many courses by themselves, such as English, Advanced Mathematics, Physics, etc. In order to reduce the blindness in taking courses, the network teaching system can intelligently recommend some suitable courses to students. How to intelligently recommend courses is the key problem to be solved. In fact, students take which course to study independently, which is closely related to their characteristics, such as majors, hobbies and genders. By Apriori algorithm in association rule mining techniques, the relationship between students' characteristics and their behavior of taking courses will be discovered. With the relations, when the first time new students take courses, the system may predict their behaviors, then recommend those courses they are interested in.

Data preparation. The data used for data mining must be clean, accurate and complete. A lot of unfiltered data may lead to poor mining results. Original data need to be filtered and cleaned before data analysis. Many students' characteristics affect results of taking courses. Here the following three characteristics are considered closely related to take courses: grade, major, gender. There are four grades: freshman, sophomore, junior, senior. The areas of major involve arts, science, engineering, medicine, and the others. Gender is male or female. A part of sample data is showed in table 1.

Table 1 the sample data set

	Grade	Classification of major	Gender	Course
1	Freshman	Engineering	Male	English
2	Sophomore	Engineering	Female	Computer technology
3	Junior	Medicine	Male	Computer technology
4	Junior	Engineering	Male	Computer technology
5	Senior	Science	Female	Advanced mathematics
6	Sophomore	Art	Male	Advanced mathematics
7	Sophomore	Engineering	Female	Advanced mathematics
8	Sophomore	Science	Male	Computer technology
9	Freshman	Science	Female	Advanced mathematics
10	Senior	Engineering	Female	Computer technology
11	Sophomore	Management	Male	English
12	Junior	Science	Female	English
13	Freshman	Medicine	Female	Advanced mathematics
14	Senior	Medicine	Male	Advanced mathematics
15	Junior	Engineering	Female	Computer technology

The grade is described as follows:A1:freshman; A2:sophomore;A3:junior;A4:senior.

The classification of major is described as follows:P1:art;P2:science;P3:engineering;P4:the others, such as medicine, management, etc.

The gender is described as follows:E1:male;E2:female.

Next, we analyze the relationship between the three students' characteristics and their behavior of taking courses.

In table 1, there are six records which show those students who take the course Computer Technology. These data are shown in Table 2. Next we use an association rule mining algorithm to get some relations between students' characteristics and their behavior of taking the course Computer Technology.

Table 2 data of the course Computer Technology in the sample data set

Grade	Classification of major	Gender
A2	P3	E2
A3	P4	E1
A3	P3	E1
A2	P2	E1
A4	P3	E2
A3	P3	E2

Association rule mining. In table2, each record corresponds to a student. There are three characteristics in each record. Each record is regarded as a transaction. Each characteristic is regarded as an item. All the data in table 2 is defined as a data set shown in Table 3.

Table 3 a transaction database

TID	Items
T01	A2,P3,E2
T02	A3,P4,E1
T03	A3,P3,E1
T04	A2,P2,E1
T05	A4,P3,E2
T06	A3,P3,E2

Next, some association rules are found out with Apriori algorithm from Table 3. The minimum Support is 33%. The itemset that at least appears in two transactions is a frequent itemset.

All of the first candidate itemsets and the Support of each itemset are listed in table4.

Table 4 the first candidate itemsets C_1 and the Support

The first candidate itemsets C_1	Count	Support[%]
A2	2	33%
A3	3	50%
A4	1	17%
P2	1	17%
P3	4	67%
P4	1	17%
E1	3	50%
E2	3	50%

From the first candidate itemsets, some itemsets are chosen as the first frequent itemsets in table5 because the Support of itemset is greater than or equal to 33%.

Table 5 the first frequent itemsets L_1 and the Support

The first frequent itemsets L_1	Count	Support [%]
A2	2	33%
A3	3	50%
P3	4	67%
E1	3	50%
E2	3	50%

According to Apriori algorithm, the second candidate itemsets generate from the results of L1⋈L1 in table6.

Table 6 the second candidate itemsets C_2 and the Support

The second candidate itemsets C_2	Count	Support [%]
(A2,P3)	1	17%
(A2,E1)	1	17%
(A2,E2)	1	17%
(A3,P3)	2	33%
(A3,E1)	2	33%
(A3,E2)	1	17%
(P3,E1)	1	17%
(P3,E2)	3	50%

Similarly, some itemsets are chosen as the second frequent itemsets in table7 from the second candidate itemsets.

Table 7 the second frequent itemsets L_2 and the Support

The second frequent itemsets L_2	Count	Support [%]
(A3,P3)	2	33%
(A3,E1)	2	33%
(P3,E2)	3	50%

Since the third candidate itemsets are not able to generate from L_2, the Apriori algorithm will stop. The last frequent itemsets include (A3,P3), (A3,E1) and (P3,E2). The conclusions can be obtained from the last frequent itemsets. If students meet any of the following conditions, they are likely to take the course Computer Technology. The first, the grade is junior and the classification of major is engineering. The second, the grade is junior and the gender is male. The third, the classification of major is engineering and the gender is male. According to these rules, the course Computer Technology will be recommended to those students who take courses the first time.

Analysis of the result. In the application above, the association rules are found out from existing sample data. The rules are the relation between the students' characteristics and the courses they take. The Apriori algorithm is executed on the data of all the students who take the same course. The output of algorithm is frequent itemsets. From the frequent itemsets, we can know which students are more likely to take this course. With these rules, when a new student first takes courses, the application system will recommend appropriate course to him/her according to his/her characteristics. But there are still several problems as follows:

Firstly, frequent itemsets are calculated on the basis of past data. The amount of sample data can't be too small. Otherwise association rules don't fit with the facts. When a large amount of new data is added to the database, we probably need to calculate frequent itemsets again and get new association rules. So the relation between students' characteristics and their tendency of taking courses is not unchanged. The relation may change with the increase of sample data. It requires that frequent itemsets are calculated repeatedly. In this way, association rules are the most consistent with the current data.

Secondly, the value of minimum Support directly affects the results of frequent itemsets. So a reasonable minimum Support is very important. According to the actual transactions, we should set an appropriate minimum Support to ensure the rationality of frequent itemsets.

Thirdly, though the repetitive implementation of Apriori algorithm can lead to more accurate results, it will take more time. This requires that the algorithm is efficient. One way is to improve the original Apriori algorithm. But another way is easier. Each time re-executing the algorithm, instead of all data, only part of the data are extracted and operated as sample data.

The above issues will be gradually improved in the future.

Conclusion

This paper introduces the theory and application of Apriori association rules mining algorithm. If students take courses to study on their own by a network teaching system, with the Apriori algorithm, we can find out the relations between students' characteristics and their behavior of

taking courses. The association rules are able to instruct new students when they first take courses. According to their characteristics, students can be recommended appropriate courses. It can increase the efficiency and reduce the blindness.

References

[1] YANGTingting, HEMingchang, OUYang, Application of Algorithm Based on Association Rule in the Higer Mathematics Exam[J]，Computer Knowledge and Technology,2013,9(30)

[2] Mehmed Kantardzic, Data Minging Concepts, Models, Methods, and Algorithms[M], IEEE Press, 2002

[3] Jiaweihan, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques(3 edition)[M], China Mechine Press,2012

[4] WANGHua, LIUPing, Application of improved association rule algorithm in early waring of student performance[J], Computer Engineering and Design,2015,36(3)

[5] FENGTao, CHANGShubao, RENYu, Combat Data Mining Based on Association Rules[J], Ship Electronic Engineering,2013,33(7)

[6] ZHUJintan, Improve of data mining Apriori algorithm[J], Electronic Design Engineering,2013,21(15)

[7] ZHAOChunling, NINGHongyun, Improvement in Apriori algorithm and using in logistics information mining[J], Journal of Tianjin University of Technology,2007,23(1)

[8] FENGXia, LIJuanjuan, YANGuan-nan, Application of association rules mining in aviation safety reports analysis[J], Computer Engineering and Design,2011,32(1)

[9] YANJie,QIWenjuan, Research Based on Aprior& FP-growth Algorithm[J],Computer Systems and Applications,2013,22[5]