A Comparative Study of Discretization Approaches for Granular Association Rule Mining

Xu He, Fan Min*, William Zhu

Lab of Granular Computing, Zhangzhou Normal University, Zhangzhou 363000, China

Abstract

Granular association rule mining is a new relational data mining approach to reveal patterns hidden in multiple tables. The current research of granular association rule mining considers only nominal data. In this paper, we study the impact of discretization approaches on mining semantically richer and stronger rules from numeric data. Specifically, the Equal Width approach and the Equal Frequency approach are adopted and compared. The setting of interval numbers is a key issue in discretization approaches, so we compare different settings through experiments on a well-known real life data set. Experimental results show that: 1) discretization is an effective preprocessing technique in mining stronger rules; 2) the Equal Frequency approach helps generating more rules than the Equal Width approach; 3) with certain settings of interval numbers, we can obtain much more rules than others.

Keywords: Granular association rule, discretization, Equal Width, Equal Frequency, relational data mining.

1. Introduction

Relational data mining schemes [7, 8] look for patterns that include multiple tables in the database. Some meaningful issues [5, 12, 10, 13, 9] are undisputed more common and more challenging than their transcriptions on a single data table. Recently, people focus on the tasks of association rule and computing with granules [14, 26, 27, 29, 33, 28].

Granular association rule mining [16, 17] is a new approach to reveal patterns hidden in multiple tables. This approach generates rules with four measures to reveal connections between concepts in two universes. We consider a database with two entities `customer` and `product` connected by a relation `buys`. An example of granular association rules might be "40% men like at least 30% kinds of alcohol; 45% customers are men and 6% products are alcohol." Here 45%, 6%,

---

40% and 30% are the *source coverage*, the *target coverage*, the *source confidence* and the *target confidence*, respectively. Numeric data are very common in real world problems. Unfortunately, only nominal data are considered in the original definition of granular association rule [16, 17].

We employed two discretization approaches, called the Equal Width approach and the Equal Frequency approach [4, 6], to preprocess the numeric data. The Equal Width approach confirms the minimum and maximum of the numeric data, and divides the range into k equal-width discrete intervals. The Equal Frequency approach confirms the minimum and maximum of the numeric data, and divides the range into k intervals which have the same number of sorted values in ascending order. Compare those two approaches by generated rules and candidates, we can obtain the strength one applied to granular association rule mining.

Experiments are undertaken on the publicly available MovieLens data set. We introduce two parameters k_1 and k_2. k_1 is the number of intervals for the age of the user, k_2 is the number of intervals for the released year of the movie. The discretization approaches are implemented with Java in our open source software COSER (Cost sensitive rough set) [22].

Our experiment results show that discretization is effective preprocessing technique in mining stronger rules. The Equal Frequency and the Equal Width approach are both simple methods to discretize data, while achieving good results. Given four measures thresholds, the Equal Frequency generates more rules than the other one. For any pair of integers (k_1, k_2), we can obtain a set of rules. Through comparing the number of all the sets of rules, we obtain certain settings of discrete interval numbers through different approaches. When setting k_1 range from 8 to 10 and k_2 range from 10 to 12 through the Equal Frequency approach, we can obtain much more rules than other settings.

The remainder of the paper is organized as follows. Section 2 reviews granular association rule. Section 3 presents granular association rules on numeric data, we might mine semantically richer and stronger rules. In Section 4, we describe each discretization approach and discuss its suitability for granular association rule mining. Experiments on the MovieLens data set [1] are discussed in Section 5. Finally, Section 6 presents the concluding remarks and further research directions.

2. Granular association rule

In this section, we revisit granular association rule [17]. We analyse the definition, and four measures of such rule. Moreover, we introduce the basic design of granular association rule mining.

2.1. The data model

First of all, we introduce the data model which is built on information systems and binary relations.

Definition 1. $S = (U, A)$ is an information system, where $U = \{x_1, x_2, \ldots, x_n\}$ is the set of all objects, $A = \{a_1, a_2, \ldots, a_m\}$ is the set of all attributes, and $a_j(x_i)$ is the value of x_i on attribute a_j for $i \in [1..n]$ and $j \in [1..m]$.

In an information system, any $A' \subseteq A$ induces an equivalence relation [23, 25]

$$E_{A'} = \{(x, y) \in U \times U | \forall a \in A', a(x) = a(y)\}, \tag{1}$$

and partitions U into a number of disjoint subsets called *blocks*. The block containing $x \in U$ is

$$E_{A'}(x) = \{y \in U | \forall a \in A', a(y) = a(x)\}. \tag{2}$$

From another viewpoint, a pair $C = (A', x)$ where $x \in U$ and $A' \subseteq A$ is called a *concept*. The *extension* of the concept is

$$ET(C) = ET(A', x) = E_{A'}(x); \tag{3}$$

while the *intension* of the concept is the conjunction of respective attribute-value pairs, i.e.,

$$IT(C) = IT(A', x) = \bigwedge_{a \in A'} \langle a : a(x) \rangle. \tag{4}$$

The *support* of the concept is the size of its extension divided by the size of the universe, namely,

$$\begin{aligned} support(C) &= support(A', x) &= support(\bigwedge_{a \in A'} \langle a : a(x) \rangle) \\ &= support(E_{A'}(x)) &= \frac{|ET(A', x)|}{|U|} \\ &= \frac{|E_{A'}(x)|}{|U|}. \end{aligned} \tag{5}$$

Definition 2. Let $U = \{x_1, x_2, \ldots, x_n\}$ and $V = \{y_1, y_2, \ldots, y_k\}$ be two sets of objects. Any $R \subseteq U \times V$ is a binary relation from U to V. The neighborhood of $x \in U$ is

$$R(x) = \{y \in V | (x, y) \in R\}. \tag{6}$$

If $U = V$ and R is an equivalence relation, $R(x)$ is the equivalence class containing x. From this definition we know immediately that for $y \in V$,

$$R^{-1}(y) = \{x \in U | (x, y) \in R\}. \tag{7}$$

A binary relation is more often stored in the database as a table with two foreign keys. In this way the storage is saved. For the convenience of illustration, here we represented it with an $n \times k$ boolean matrix.

With Definitions 1 and 2, we propose the following definition.

Definition 3. [16] A many-to-many entity-relationship system (MMER) is a 5-tuple $ES = (U, A, V, B, R)$, where (U, A) and (V, B) are two information systems, and $R \subseteq U \times V$ is a binary relation from U to V.

2.2. Granular association rule with four measures

Now we come to the central definition of granular association rules.

Definition 4. [16] A *granular association rule* is an implication of the form

$$(GR) : \bigwedge_{a \in A'} \langle a : a(x) \rangle \Rightarrow \bigwedge_{b \in B'} \langle b : b(y) \rangle, \tag{8}$$

where $A' \subseteq A$ and $B' \subseteq B$.

According to Equation (5), the set of objects meeting the left-hand side of the granular association rule is

$$LH(GR) = E_{A'}(x); \tag{9}$$

while the set of objects meeting the right-hand side of the granular association rule is

$$RH(GR) = E_{B'}(y). \tag{10}$$

The *source coverage* of a granular association rule is

$$scoverage(GR) = |LH(GR)|/|U|. \tag{11}$$

The *target coverage* of GR is

$$tcoverage(GR) = |RH(GR)|/|V|. \tag{12}$$

There is a tradeoff between the source confidence and the target confidence of a rule. Consequently, no values can be obtained directly from the rule. To compute any one of them, we should specify the threshold of the other. Let tc be the target confidence threshold. The *source confidence* of the rule is

$$sconfidence(GR, tc) = \frac{|\{x \in LH(GR)| \frac{|R(x) \cap RH(GR)|}{|RH(GR)|} \geq tc\}|}{|LH(GR)|}. \tag{13}$$

Let mc be the source confidence threshold, and

$$\begin{aligned}
&|\{x \in LH(GR)||R(x) \cap RH(GR)| \geq K+1\}| \\
&< mc \times |LH(GR)| \\
&\leq |\{x \in LH(GR)||R(x) \cap RH(GR)| \geq K\}|.
\end{aligned} \tag{14}$$

This equation means that $mc \times 100\%$ elements in $LH(GR)$ have connections with at least K elements in $RH(GR)$, but less than $mc \times 100\%$ elements in $LH(GR)$ have connections with at least $K+1$ elements in $RH(GR)$. The *target confidence* of the rule is

$$tconfidence(GR, mc) = K/|RH(GR)|. \tag{15}$$

In fact, the computation of K is non-trivial. First, for any $x \in LH(GR)$, we need to compute $tc(x) = |R(x) \cap RH(GR)|$ and obtain an array of integers. Second, we sort the array in a descending order. Third, let $k = \lfloor mc \times |LH(GR)| \rfloor$, K is the k-th element in the array.

2.3. Granular association rule mining

The basic design of granular association rule mining is as follows.

Definition 5. The granular association rule mining.

Input: An $ES = (U, A, V, B, R)$, a minimal source coverage threshold ms, a minimal target coverage threshold mt, a minimal source confidence threshold mc, and a minimal target confidence threshold tc.

Output: All granular association rules satisfying $scoverage(GR) \geq ms$, $tcoverage(GR) \geq mt$, $sconfidence(GR) \geq mc$, and $tconfidence(GR) \geq tc$.

3. Granular association rule on numeric data

There are many different types of data to describe objects. Recently, all data are implicitly considered to be nominal. However, in the real world applications, a very large proportion of data sets involve numerical data. One scheme to solve this problem is to divide numeric data into a number of intervals and regard each interval as a category. This process is usually named discrerization [**?** 3, 15, 18, 15, 18]. At present, the most important thing we intend to do is that we can mine semantically richer and stronger rules which cannot mine in primary data through discretization. For instance, we give an information system in Table 1, where $U = \{$c1, c2, c3, c4, c5, c6, c7, c8, c9, c10$\}$, and A = {Age, Gender, Married, Salary}. Among them, Age and Salary values are numeric data. Another example is given by Table 2, where $U = \{$p1, p2, p3, p4, p5, p6, p7, p8$\}$, and A = {Country, Category, Color, Price}. Among them, Price values are numeric data.

A binary relation is more often stored in the database as a table with two foreign keys. In this way the storage is saved. For the convenience of illustration, here we represented it with an $n \times k$ boolean matrix. An example is given by Table 3, where U is the set of customers as indicated by Table 1, and V is the set of products as indicated by Table 2.

At present, we indicate all of the numeric data from the information systems. And then divide numeric data into a number of intervals and regard each interval as a category, as shown in Tables 4, 5. From the MMER given by Tables 3, 4and 5 we may obtain the following interesting rule.

(Rule 1) ⟨Gender: Male⟩ $\Rightarrow$ ⟨Category: Alcohol⟩.

(Rule 2) ⟨Age: [30, 35)⟩ $\wedge$ ⟨Gender: Male⟩ $\Rightarrow$ ⟨Category: Alcohol⟩.

(Rule 3) ⟨Married: Yes⟩ $\Rightarrow$ ⟨Country: China⟩.

(Rule 4) ⟨Married: Yes⟩ $\wedge$ ⟨Salary: [4700, 5600]⟩

   $\Rightarrow$ ⟨Country: China⟩ $\wedge$ ⟨Price: [2.0, 7.3)⟩.

Rule 1 can be read as "men like alcohol." Rule 2 can be read as "men whose age is between 30 and 35 like alcohol." Rule 3 can be read as "Married people like products made in China." Rule 4 can be read as "Married people whose salaries are between 4700 and 5600, like products made in China, which prices are between 2.0 and 7.3."

From above we can come to a conclusion, we can mine semantically richer and stronger rules which cannot be mined in primary data through discretization,

Table 1: Customer

CID	Age	Gender	Married	Salary
c1	20	Male	No	2000
c2	25	Female	Yes	2800
c3	23	Male	No	3500
c4	26	Female	Yes	2400
c5	32	Male	Yes	5600
c6	36	Male	Yes	4200
c7	39	Male	Yes	5000
c8	40	Female	Yes	5000
c9	35	Female	Yes	3400
c10	34	Male	Yes	3600

Table 2: Product

PID	Country	Category	Color	Price
p1	China	Staple	Yellow	2.0
p2	Australia	Staple	Black	4.0
p3	China	Daily	White	5.5
p4	China	Meat	Red	8.0
p5	Australia	Meat	Red	18.0
p6	China	Alcohol	Yellow	3.0
p7	France	Alcohol	Yellow	5.0
p8	France	Alcohol	White	16.5

such as Rules 2, 4. Given the same four measures threshold, Rule 2 has a semantically richer rule than Rule 1, and Rule 4 has a richer rule than Rule 3. A detailed explanation of Rule 4 might be "60% married people like at least 60% products, which prices are between 2.0 and 7.3; 70% customers are married people, 62.5% products of all products which prices are between 2.0 and 7.3."

4. Discretization approaches

In this section, we introduce different discretization approaches, which can divide the numeric data into different intervals and regard each interval as a category. Given four measures thresholds, we can mine different rules. Since the number of intervals is a key issue in discretization approaches, we try to use some different settings of interval numbers to can obtain the suitable one. Then we can mine appropriate granule association rules.

In this paper, we adopt two discretization approaches, namely the Equal Width approach and the Equal Frequency approach. The two approaches are

Table 3: Buys

CID\ PID	p1	p2	p3	p4	p5	p6	p7	p8
c1	1	0	0	1	1	1	0	0
c2	1	0	0	1	0	1	0	0
c3	0	0	1	0	1	0	1	1
c4	0	1	0	1	1	1	0	0
c5	0	1	1	1	0	0	1	1
c6	0	1	0	1	0	0	1	0
c7	1	1	1	1	0	0	1	1
c8	0	1	1	0	1	1	1	0
c9	1	0	1	0	1	0	1	0
c10	1	0	1	0	1	0	1	1

Table 4: Discretization for Age and Salary

CID	Age	Gender	Married	Salary
c1	[20,25)	Male	No	[2000, 2900)
c2	[25,30)	Female	No	[2000, 2900)
c3	[20,25)	Male	No	[2900, 3800)
c4	[25,30)	Female	Yes	[2000, 2900)
c5	[30,35)	Male	Yes	[3800, 4700]
c6	[35,40]	Male	Yes	[2900, 3800)
c7	[35,40]	Male	Yes	[4700, 5600)
c8	[35,40]	Female	Yes	[4700, 5600)
c9	[35,40]	Female	Yes	[2900, 3800)
c10	[30,35)	Male	Yes	[2900, 3800)

both simple methods to discretize data and have often been used to produce nominal data from numeric ones.

4.1. The Equal Width approach

The Equal Width approach confirms the minimal value a_0 and the maximal value a_k of the numeric data, and divides the range into k equal-width discrete intervals. Here k is a parameter supplied by the user. The approach calculates the discretization width

$$\lambda = \frac{a_k - a_0}{k}. \tag{16}$$

These values form the boundary set $\{a_0, a_1, ..., a_i, ..., a_{k-1}, a_k\}$ for $\{[a_0, a_1), ..., [a_{i-1}, a_i), ..., [a_{k-1}, a_k]\}$, $a_i = a_0 + i\lambda$, where $i = 1, 2, ..., k$. The approach is applied to each numeric data independently. Finally, we obtain discretization data.

Table 5: Discretization for Price

PID	Country	Category	Color	Price
p1	China	Staple	Yellow	[2.0, 7.3)
p2	Australia	Staple	Black	[2.0, 7.3)
p3	China	Daily	White	[2.0, 7.3)
p4	China	Meat	Red	[7.3, 12.7)
p5	Australia	Meat	Red	[12.7, 18.0]
p6	China	Alcohol	Yellow	[2.0, 7.3)
p7	France	Alcohol	Yellow	[2.0, 7.3)
p8	France	Alcohol	White	[12.7, 18.0]

4.2. The Equal Frequency approach

The Equal Frequency approach confirms the minimal value b_0, the maximal value b_k of the numeric data, and sorts the values from in ascending order. Here k is a parameter supplied by the user. Divide the range into k of intervals in order that every interval involves the same number of sorted values, These values form the boundary set $\{b_0, b_1, b_2, ..., b_{k-1}, b_k\}$ for $\{[b_0, b_1), [b_1, b_2), ..., [b_{k-1}, b_k]\}$.

We set different interval number k to divide the numeric data, and use different discretization approaches to produce different intervals. We know that more interval numbers, higher confidence of intervals, and lower coverage of intervals. Compare those intervals to get the suitable one for rule mining. For example, Table 2 shows that the value of Price range from 2.0 to 18.0. Set $k = 3$, we get the price of p3 is between 2.0 and 7.3 with the Equal Width approach, while it is between 2.0 and 7.0 with the Equal Frequency approach. Set $k = 4$, we get the price of p3 is between 2.0 and 6.0 with the Equal Width approach, while it is between 2.0 and 5.5 with the Equal Frequency approach. Comparing those intervals, we obtain that take advantage of interval numbers and discretization approach is very important to produce suitable intervals for mining rule.

5. Experiments on a real world data set

5.1. A movie rating data set

The MovieLens data set [1] assembled by the GroupLens project is widely used in recommender systems (see, e.g., [2, 11, 24, 20, 21]). We downloaded the data set from the Internet Movie Database [1]. The data set contains 100,000 ratings (1-5) from 943 users on 1,682 movies, with each user rating at least 20 movies [24]. In order to run our algorithm, we preprocessed the data set as follows.

1. Remove movie names. They are not useful in generating meaningful granular association rules.

2. Use release year instead of release date. In this way the granule is more suitable.

3. Select the movie genre. In the original data, the movie genre is multi-valued since one movie may fall in more than one genre. For example, a movie can be both Animation and Children's. Unfortunately, granular association rules do not support this type of data at this time. Since the main objective of this work is to test compare the performances of algorithms, we use a simple approach to deal with this issue. That is to sort movie genres according to the number of users they attract, and only keep the one highest priority for the current movie. We adopt the following priority (from high to low): Comedy, Action, Thriller, Romance, Adventure, Children, Crime, Sci-Fi, Horror, War, Mystery, Musical, Documentary, Animation, Western, FilmNoir, Fantasy, Unkown.

Our database schema is as follows.

- User (<u>userID</u>, age, gender, occupation)
- Movie (<u>movieID</u>, releaseYear, genre)
- Rates (<u>userID, movieID</u>)

According to given intervals $[0, 18)$, $[18, 25)$, $[25, 30)$, $[30, 35)$, $[35, 45)$, $[45, 56)$, $[56, \infty)$, the age of the user is discretized by the GroupLens project. And then we use release decade instead of release date for the movies range from 1920s to 1990s. As a result, a manual discretization setting is given to divide numeric data to obtain a finer granule. The setting would be used to compare with other discretization approaches.

5.2. Results

In this section, we try to answer the following problems through experimentation.

1. Compared with the manual discretization setting to mine rules, Which approach outperform, the Equal Width approach or the Equal Frequency approach?

2. Whether we can mine much semantically richer rules through discretization?

3. What are the certain settings of discrete interval numbers for the numeric data?

We undertake three sets of experiments to answer the questions one by one.

5.2.1. The performance of discretization approaches

The evaluation of discretization approaches was performed using the number of generated rules and candidates. We compare the Equal Width approach, the Equal Frequency approach, the manual discretization setting and primary data which is without discretization. Let $mc = 0.15$, $tc = 0.17$ and $ms = mt \in \{0.04, 0.06, 0.08, 0.10, 0.12\}$. Suppose k is the number of intervals. We set $k = 4$ and
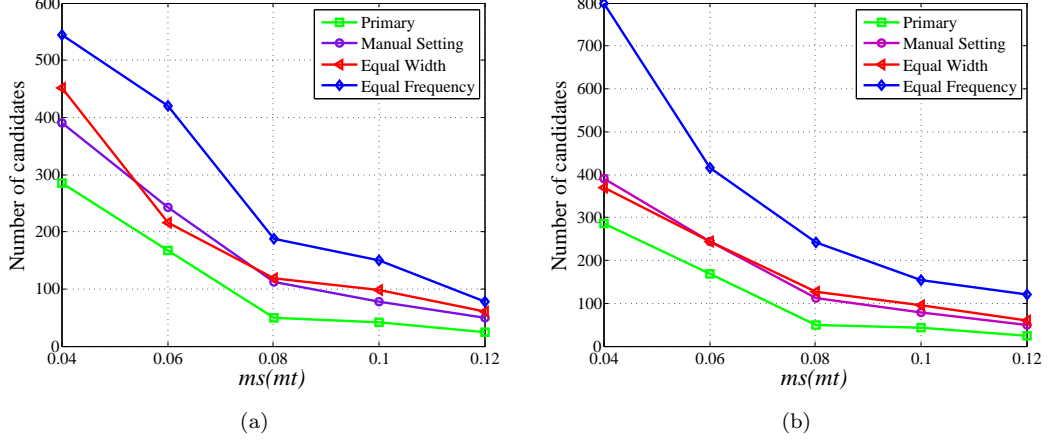
Figure 1: Number of candidates: (a) interval number $k = 4$; (b) interval number $k = 8$.

$k = 8$ for rule mining, respectively. We compare the number of candidates and rules, as shown in Figures 1, 2.

Figures 1, 2 show that all discrete approaches can help to mine more candidates and rules from discreted data than not do it from primary data, and the Equal Frequency mine the most. When $ms = mt = 0.12$, the Equal Frequency can still mine rules, but the others cannot mine any.

We compare the Equal Width approach and the manual discretization setting. When $k = 4$, the number of candidates and rules of the Equal Width approach and the manual discretization setting have big different, the reason is that a interval may divide into some intervals, which have affects on the number of generated rules. For example, the Equal Width approach obtain a interval [1979, 1998], which includes 1980s and 1990s. Specifically, when $k = 8$, the number of candidates and rules of them is very similar, the reason is that each interval of them is very similar.

5.2.2. The semantically richer

We obtain some strong rules using Equal Width and Equal Frequency. Here we set interval number $k = 4$, $ms = mt = 0.06$, $mc = 0.15$, and $tc = 0.17$. 43 and 68 granular association rules are respectively obtained by Equal Width and Equal Frequency. We respectively list 4 rules of them below.

The Equal Width approach:

(Rule 6) ⟨age [7,24]⟩
    ⇒ ⟨genre: action⟩

(Rule 7) ⟨age [7,24]⟩ ∧ ⟨gender: male⟩
    ⇒ ⟨genre: action⟩

(Rule 8) ⟨age [7,24]⟩ ∧ ⟨gender: male⟩
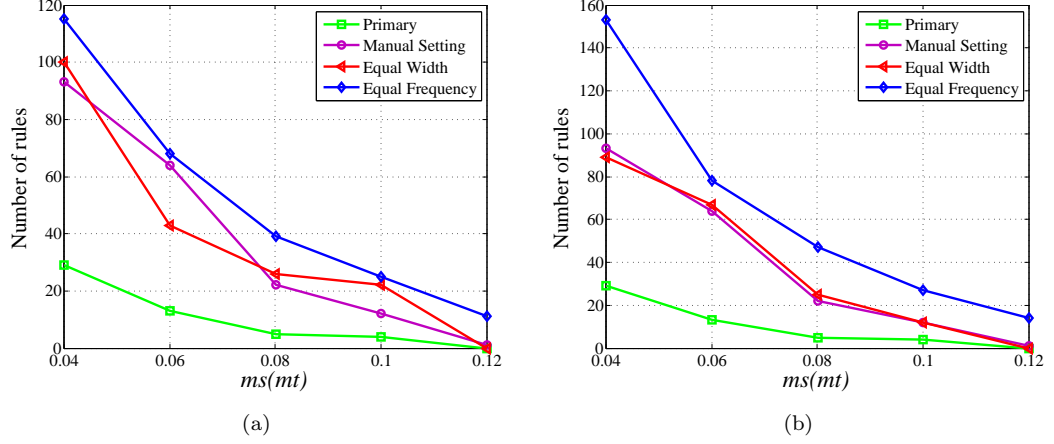   ⇒ ⟨releaseYear: [1979,1998]⟩ ∧ ⟨genre: action⟩

10

Figure 2: Number of rules: (a) interval number $k = 4$; (b) interval number $k = 8$.

(Rule 9) $\langle$age [7,24)$\rangle \wedge \langle$gender: male$\rangle \wedge \langle$occupation: student$\rangle$
     $\Rightarrow \langle$releaseYear: [1979,1998]$\rangle \wedge \langle$genre: action$\rangle$
The Equal Frequency approach:
(Rule 10) $\langle$age [7,25)$\rangle$
     $\Rightarrow \langle$genre: action$\rangle$
(Rule 11) $\langle$age [7,25)$\rangle \wedge \langle$occupation: student$\rangle$
     $\Rightarrow \langle$genre: action$\rangle$
(Rule 12) $\langle$age [25,31)$\rangle \wedge \langle$gender: male$\rangle$
     $\Rightarrow \langle$releaseYear: [1992,1995]$\rangle \wedge \langle$genre: comedy$\rangle$
(Rule 13) $\langle$age [7,25)$\rangle \wedge \langle$gender: male$\rangle \wedge \langle$occupation: student$\rangle$
     $\Rightarrow \langle$releaseYear: [1992,1995]$\rangle \wedge \langle$genre: comedy$\rangle$

All rules are quite meaningful from different discrete approaches, and they might be applied to movie recommendation directly. For Rule 6 indicates that user whose age range from 7 to 24 rate action movies. We observe that Rule 7 and Rule 8 is finer than Rule 6, which is in turn semantically richer than Rule 6. Rule 9 obtains the semantically richest rule. For Rule 11 indicates that user whose age range from 7 to 25 rate action movies. We observe that Rule 11 is finer than Rule 10, it is similar to the above. Rule 12 mine user age range 25 to 31, but not range 7 to 25, and Rule 13 mine movie genre is comedy but not action, those rules cannot be comparable with Rule 11, but still useful.

5.2.3. The setting of interval numbers

The setting of interval numbers is a key issue in discretization approaches, so we compare different settings through experiment. We introduce two parameters k_1, k_2, k_1 is number of interval for the numeric data of *User*, k_2 is number of interval for the numeric data of *Movie*.

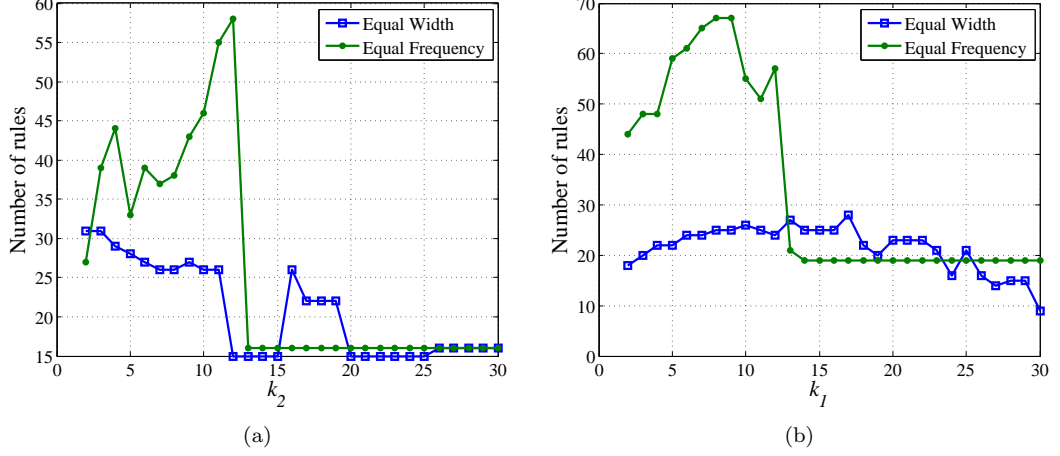We set $ms = mt = 0.08$, $mc = 0.15$, and $tc = 0.17$. Firstly we let $k_1 = 10$

11

Figure 3: Number of rules: (a) $k_1 = 10$; (b) $k_2 = 11$.

and let k_2 increases from 2 to 30, the number of rules are compared, as depicted in Figure 3(a). Secondly we let $k_2 = 11$ and let k_1 increases from 2 to 30, the number of rules are compared, as drew in Figure 3(b). Thirdly we let k_1, k_2 increase from 2 to 20, respectively, and obtain the corresponding to number of rules, we draw a three-dimensional figure, as shown in Figures 4 and 5.

Figure 3(a) shows the number of rules decreases as k_2 increases, the reason is more interval numbers and lower coverage of intervals, some rules do not satisfy mt that we cannot mine them. The Equal Frequency approach can mine much more rules than the Equal Width approach at begin. This is because the number of the users and the movies are well-distributed in the intervals divided by the Equal Frequency approach, more and more intervals can satisfy mt that we can mine much more rules. When $k_2 = 12$ of Equal Width and $k_2 = 13$ of Equal Frequency, the number of rules slumps, the reason is some rules do not satisfy mt. For example, when $k_2 = 12$, the number of candidates is 18×15, while $k_2 = 13$, the number of candidates is only 18×3, which is much less. Finally, the number of rules remains unchanged, because only these rules can be mined before $k_2 = 30$.

Figure 3(b) also shows the number of rules decreases as k_1 increases, this is because more interval numbers and lower coverage of intervals, some rules do not satisfy ms that we cannot mine them. The Equal Frequency approach can mine much more rules when k_1 is between 2 and 12. For the Equal Frequency approach, when $k_1 = 13$, the number of rules slumps. This is because some rules do not satisfy ms. For instance, when $k_2 = 12$, the number of candidates is 19×14, while $k_2 = 13$, the number of candidates is only 8×14, which is much less. Between $k_1 = 14$ and $k_1 = 30$, the number of rules remains unchanged, this is because only these rules can be mined. For the Equal Width approach, it decreases stable as k_1 increases.
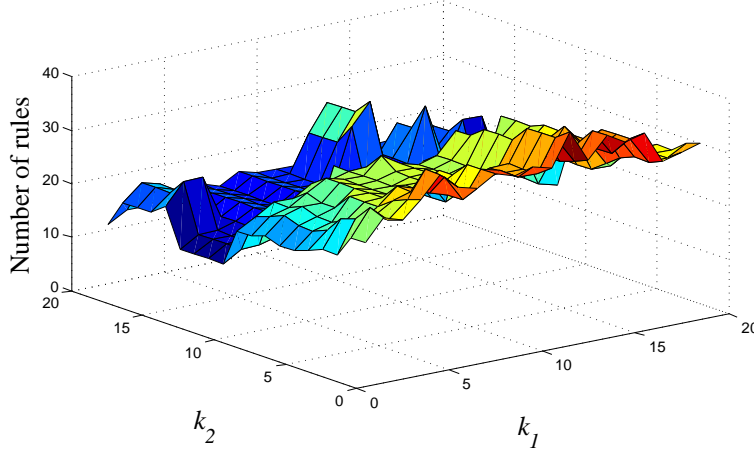
12

Figure 4: Different settings of interval numbers obtain number of rules through the Equal Width approach

Figures 4 and 5 indicate the number of rules changes with k_1 and k_2 increase. The Equal Frequency approach can mine more rules than Equal Width. For Figure 4, while k_1 range from 10 to 13 and k_2 range from 9 to 11, we can obtain more rules. For Figure 5, while k_1 range from 8 to 10 and k_2 range from 10 to 12, we can obtain more rules. Compare those two Figures, we observe that Figure 5 is more intuitive than Figure 4.

5.3. Discussions

Now we can answer the questions proposed at the beginning of this section.

1. Discretization is an effective preprocessing technique in mining stronger rules, so it outperforms the primary data. Compared with the manual discretization setting to mine rules and the Equal Width approach, the Equal Frequency approach generates more candidates number and stronger rules.
2. Through discretization, we can obtain much semantically richer rules.
3. When setting k_1 range from 8 to 10 and k_2 range from 10 to 12 for Equal Frequency, we obtain certain settings of discrete interval numbers.

6. Conclusions and further works

In this paper, we introduced an evaluation and comparison of discretization approaches for granular association rule mining. With the help of discretization, we mined semantically richer and stronger rules. The Equal Frequency approach helped generating more rules than the Equal Width approach. We obtained certain settings of discrete interval to mine much more rules through different approaches.

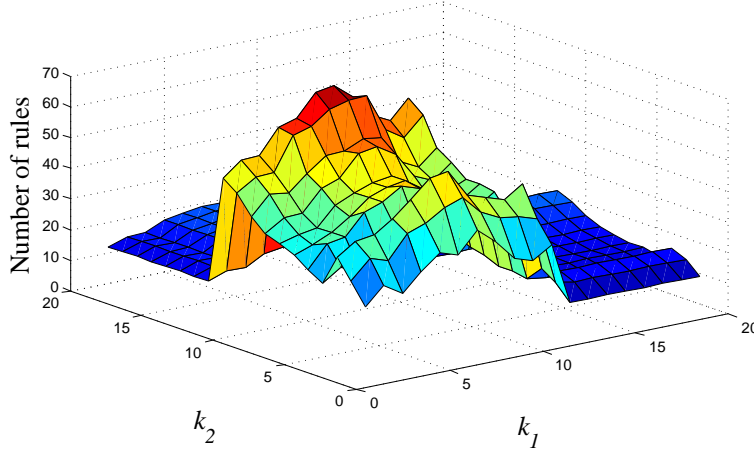The following research topics deserve further investigation:

13

Figure 5: Different settings of interval numbers obtain number of rules through the Equal Frequency approach

1. Preferable discretization approaches. In this work we adopt the Equal Width approach and the Equal Frequency approach. In fact, there are a lot of discretization approaches. Many approaches such as rough sets and decision trees would work better on discretized data [30, 31, 32, 19]. We will try to choose some suitable discretization approaches, and design a more appropriate one for granular association rule mining.

2. Intelligent choice. In practice, some data sets contain different numeric data of different attributes, and we use the same discretization approaches to deal with them. However, different algorithms adapt to different data, so that we try to group different algorithms to realize intelligent choice for discretization of the same data set. The improved scheme is more valuable in practical application.

Acknowledgements

References

[1] Internet movie database.
URL http://movielens.umn.edu

[2] M. Balabanović, Y. Shoham, Fab: content-based, collaborative recommendation, Communication of ACM 40 (3) (1997) 66–72.

14

[3] S. Bay, Multivariate discretization of continuous variables for set mining, in: Proceedings of the 6th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2000.

[4] D. Chiu, A. Wong, B. Cheung, Information discovery through hierarchical maximum entropy discretization and synthesis, G. Piatetsky-Shapiro, W.J. Frawley (Eds.), Knowledge Discovery in Databases, MIT Press, Cambridge, Mass (1991) 125–140.

[5] L. Dehaspe, H. Toivonen, R. D. King, Finding frequent substructures in chemical compounds, in: 4th International Conference on Knowledge Discovery and Data Mining, 1998.

[6] J. Dougherty, R. Kohavi, M. Sahami, Supervised and unsupervised discretization of continuous features, in: Machine learning international workshop then conference, Morgan Kaufmann Publishers, Inc., 1995.

[7] S. Džeroski, Multi-relational data mining: An introduction, in: SIGKDD Explorations, vol. 5, 2003.

[8] S. Džeroski, N. Lavrac (eds.), Relational data mining, Springer, 2001.

[9] B. Goethals, W. L. Page, M. Mampaey, Mining interesting sets and rules in relational databases, in: Proceedings of the 2010 ACM Symposium on Applied Computing, 2010.

[10] B. Goethals, W. L. Page, H. Mannila, Mining association rules of simple conjunctive queries, in: Proceedings of the SIAM International Conference on Data Mining (SDM), 2008.

[11] J. L. Herlocker, J. A. Konstan, A. Borchers, J. Riedl, An algorithmic framework for performing collaborative filtering, in: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '99, 1999.

[12] V. C. Jensen, N. Soparkar, Frequent itemset counting across multiple tables, in: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Application, vol. 1805 of LNCS, 2000.

[13] Y. Kavurucu, P. Senkul, I. Toroslu, ILP-based concept discovery in multi-relational data mining, Expert Systems with Applications 36 (2009) 11418–11428.

[14] T. Y. Lin, Granular computing on binary relations i: Data mining and neighborhood systems, in: Rough Sets in Knowledge Discovery, 1998.

[15] F. Min, H. B. Cai, Q. H. Liu, Z. J. Bai, Dynamic discretization: a combination approach, in: Proceedings of International Conference on Machine Learning and Cybernetics, 2007.

[16] F. Min, Q. H. Hu, W. Zhu, Granular association rules on two universes with four measures, submitted to Information Sciences.
URL http://arxiv.org/abs/1209.5598

[17] F. Min, Q. H. Hu, W. Zhu, Granular association rules with four subtypes, in: Proceedings of the 2011 IEEE International Conference on Granular Computing, 2012.

[18] F. Min, Q. Liu, C. Fang, Rough sets approach to symbolic value partition, International Journal of Approximate Reasoning 49 (2008) 689–700.

[19] F. Min, W. Zhu, A competition strategy to cost-sensitive decision trees, in: Proceedings of Rough Set and Knowledge Technology, vol. 7414 of LNAI, 2012.

[20] F. Min, W. Zhu, Granular association rule mining through parametric rough sets, in: Proceedings of the 2012 International Conference on Brain Informatics, vol. 7670 of LNCS, 2012.

[21] F. Min, W. Zhu, Granular association rule mining through parametric rough sets for cold start recommendation.
URL http://arxiv.org/abs/1210.0065

[22] F. Min, W. Zhu, H. Zhao, G. Y. Pan, J. B. Liu, Z. L. Xu, X. He, Coser: Cost-senstive rough sets, http://grc.fjzs.edu.cn/~fmin/coser/ (2012).

[23] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (1982) 341–356.

[24] A. I. Schein, A. Popescul, L. H. Ungar, D. M. Pennock, Methods and metrics for cold-start recommendations, in: SIGIR '02, 2002.

[25] A. Skowron, J. Stepaniuk, Approximation of relations, in: W. Ziarko (ed.), Proceedings of Rough Sets, Fuzzy Sets and Knowledge Discovery, 1994.

[26] J. T. Yao, Y. Y. Yao, Information granulation for web based information retrieval support systems, in: Proceedings of SPIE, vol. 5098, 2003.

[27] Y. Y. Yao, Granular computing: basic issues and possible solutions, in: Proceedings of the 5th Joint Conference on Information Sciences, vol. 1, 2000.

[28] Y. Y. Yao, A partition model of granular computing, Transactions on Rough Sets I (2004) 232–253.

[29] L. Zadeh, Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems 19 (1997) 111–127.

[30] W. Zhu, Generalized rough sets based on relations, Information Sciences 177 (22) (2007) 4997–5011.

[31] W. Zhu, Relationship among basic concepts in covering-based rough sets, Information Sciences 17 (14) (2009) 2478–2486.

[32] W. Zhu, Relationship between generalized rough sets based on binary relation and covering, Information Sciences 179 (3) (2009) 210–225.

[33] W. Zhu, F. Wang, Reduction and axiomization of covering generalized rough sets, Information Sciences 152 (1) (2003) 217–230.