A Quantitative Data Mining Algorithm Based on Improved Multi-level Fuzzy Association Rules (Postprint)

Authors: Zhang Dingxiang, Yuejin Zhang

Date: 2018-10-11T00:00:00+00:00

Abstract

To address the problems of low rule extraction accuracy, long algorithm runtime, and difficulty in meeting user requirements associated with single-level structure rule extraction, this paper proposes a quantitative data mining algorithm based on improved multi-level fuzzy association rules. By employing high-frequency itemsets and forming a top-down mining process through continuously deepening iteration, the algorithm integrates fuzzy set theory, data mining algorithms, and multi-level classification techniques to discover fuzzy association rules from transaction datasets, mine the implicit knowledge of quantitative value information stored in multi-level structured transaction databases, and fulfill users' customized information mining requirements. Experimental results demonstrate that the proposed data mining algorithm exhibits significant advantages in both mining accuracy and computational time compared with other algorithms, which can bring breakthrough progress to the practical application of multi-level association rule extraction methods.

Full Text

Preamble

Quantitative Data Mining Algorithm Based on Improved Multi-Level Fuzzy Association Rules

Zhang Dingxiang[1], Zhang Yuejin[2]

(1. College of Computer & Information Engineering, Guizhou University of Commerce, Guiyang 550014, China;
2. School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: Existing rule extraction methods based on single-level hierarchical structures suffer from low accuracy, long running times, and difficulty meeting user requirements. To address these issues, this paper proposes a quantitative data mining algorithm based on improved multi-level fuzzy association rules. The algorithm employs high-frequency itemsets and forms a top-down mining process through progressively deepening iterations. By integrating fuzzy set theory, data mining algorithms, and multi-level classification techniques, it discovers fuzzy association rules from transaction datasets, excavates hidden knowledge of quantitative value information stored in multi-level structured transaction databases, and fulfills users' customized information mining needs. Experimental results demonstrate that the proposed data mining algorithm offers significant advantages in mining precision and computational time compared to other algorithms, potentially bringing breakthrough progress to practical applications of multi-level association rule extraction methods.

Keywords: fuzzy set; user customization; multi-level structure; flexible boundary; membership function

0 Introduction

In recent years, as theoretical frameworks and algorithms in data science have matured, scientific research based on data algorithms has gradually become a focal point in both academia and industry. Among these, data mining theory has emerged as a key research area in data science for extracting relational information. Based on the type of data information being mined, data mining methods can be further categorized into association mining, classification mining, clustering mining, and sequential mining. Association mining represents an important type of data mining that primarily identifies correlations among different items in transaction databases. Association mining methods have been widely applied in market planning and marketing strategy development, yielding favorable results. For instance, supermarket managers can use association mining to effectively predict which product combinations customers are more likely to purchase together. Classic rules such as "customers who buy diapers usually also buy beer" can be discovered through this process. Based on these association rules, managers can place beer and diapers in close proximity to encourage simultaneous purchases. Evidently, research on quantitative data mining of association rules holds significant importance.

Surveying recent academic research on association mining algorithms reveals that most are based on the Apriori algorithm, which gradually generates and tests candidate itemsets. However, this process typically requires repeated scans of the database, incurring high computational costs. As the sample size for association mining grows exponentially, the high time cost consumed by the Apriori algorithm has become a critical issue demanding urgent resolution in association mining research. Consequently, literature [7] proposed that associa-

tion rules should satisfy two user-specific constraints—confidence and support—to reduce computational time in data mining. Support is defined as the proportion of transactions in the transaction set that satisfy a given condition, while confidence is defined as the ratio of the support of transactions satisfying the condition to the support of the transaction set.

Furthermore, the vast majority of current association rule algorithm research focuses solely on single-concept-level mining, with less attention given to multi-concept-level mining, as seen in the algorithms of literature [8] and [9]. Literature [10] proposed a fuzzy mining algorithm that applies multi-level association mining to provide users with more valuable information from a practical application perspective. However, when providing solutions for redundant rules, the iterative algorithm becomes complex and consumes substantial computational resources.

In response to the aforementioned research status, this paper proposes a mining algorithm based on improved multi-level fuzzy association rules for extracting implicit information from quantitative data. This method employs high-frequency itemsets and forms a top-down mining process through progressively deepening iterations. The algorithm integrates fuzzy set theory, data mining algorithms, and multi-level classification techniques to discover fuzzy association rules from transaction datasets. Experimental results combined with specific examples verify the algorithm' s superiority. For users, the rules mined by this method are more logical and better aligned with human cognitive patterns.

---

1 Proposed Improved Multi-Level Fuzzy Association Rule Mining Algorithm

To mine association rules across multiple concept levels, it is necessary to classify items or effectively define conceptual hierarchical structures. Conceptual hierarchies can be derived from a directed acyclic graph (DAG). A conceptual hierarchy represents the relationship between item paths and requirements, enabling their classification at different abstraction levels. These concept hierarchies are either readily available or can be obtained through domain expert application. For example, a user is typically not only interested in the association between computers and printers but 更希望获得台式电脑的价格与激光打印机的价格之间的关联。此外，模糊理论 [13] 对于多层次关联挖掘方法的研究具有一定借鉴意义，这一理论提出通过引入渐进成员关系来表征语言术语的模糊边界 [14]。

因此，为实现定量数据集中多层次关联规则的有效挖掘，将基于分类学理论成果 [15]，提出一种改进的模糊挖掘算法。这一算法综合利用数据挖掘方法、多层次分类理论以及隶属函数定义，可用于在给定的事务数据集中挖掘模糊关联规则。

1.1 Improved Multi-Level Association Rules

Mining association rules at multiple concept levels may yield more generalizable and applicable rules. In practical application scenarios, the classification of specific items is typically predefined and can be represented using a structure tree. Terminal nodes of the structure tree represent actual items appearing in transactions; internal nodes represent concepts formed by lower-level nodes.

In Figure 1, the root node is at level 0, internal nodes representing categories (such as beverages) are at level 1, internal nodes representing flavors (such as lemon) are at level 2, and terminal nodes representing brands (such as Coca-Cola) are at level 3. Ultimately, only terminal nodes appear in transactions during the algorithm process. In the predefined classification, nodes are first encoded as combinations of numbers and the symbol "*" based on their positions in the structure tree. For example, in Figure 1, the internal node "juice" is encoded as 1**, the internal node "strawberry flavor" is encoded as 11*, and the terminal node "Huiyuan" is encoded as 111.

[Figure 1: see original paper] Schematic diagram of structure tree encoding based on beverage classification

Using the encoding rule provided by formula (1), any node in a structural hierarchy can be encoded. To facilitate demonstration of the specific encoding process operations, Figure 2 shows the encoding of each node using a typical four-level structure as an example. After encoding is completed, each item in the transaction database is replaced with its corresponding code.

[Figure 2: see original paper] Example diagram of a typical four-level structure encoding

1.2 Establishment of Multi-Level Fuzzy Association Mining Algorithm

The algorithm establishment steps for the multi-level fuzzy association mining algorithm are as follows:

A symbol sequence composed of the dataset and the symbol "*" is used to encode predefined groupings, which can be obtained according to formula (1):

$$10jDS = \times + 01 * *N * *11*$$

where: j is the position index of the node at the current level (node position indices are consecutive integers starting from 1, with each node encoded sequentially from left to right); D is the encoding of the node at the current level; S is the encoding of the parent node at the current level.

For each transaction data item, where i represents the transaction index with an upper limit equal to the total number of transactions in the database, items with identical first f bits are aggregated together to calculate their support.

Groups with support less than the minimum support threshold at the current level are removed.

For different data items, distinct membership functions are predefined to characterize the differences among various items. Each data item possesses unique attributes and membership functions. Subsequently, the grouped values of each transaction data item are transformed into fuzzy sets through mapping by specific membership functions. The transformation formula is shown in equation (2):

$$\frac{1}{)(hssabcbhcfR} = \sum salZsalSumsalZ$$

All transactions in the transaction dataset are combined and partitioned into fuzzy sets using the method in equation (2). The value of each fuzzy region in the fifteen data points is calculated according to equation (3), where the sum of all values equals the total.

Then, according to equation (4), the region is designated. If the region' s value at the current level is greater than or equal to the minimum support (K), it is placed in the frequent 1-itemset.

For different hierarchical index values, the following processes are executed:

a) If a candidate set is generated in the second hierarchical structure, where the set represents multiple candidate items at level k, this indicates the algorithm can continue to be applied. These items are frequent items obtained through fuzzy level crossing methods from various levels. For example, candidate 2-itemsets at level 2 are not limited to frequent item pairs at level 2; frequent items at level 2 may also combine with frequent items at level 1 to form candidate 2-itemsets at level 2. However, according to basic classification algorithm theory, each 2-itemset in the candidate itemset must contain at least one item from the set, and the next item must not be a taxonomic ancestor of that item. All possible 2-itemsets are collected. After obtaining this collection, step b) is executed.

b) If the hierarchical structure index > 2, the candidate set needs to be generated through software methods. This is a candidate itemset with multiple items at level k generated from the previous level, using a method similar to the Apriori algorithm' s candidate generation.

For any candidate r-itemset obtained through screening in the set:

a) Calculate the fuzzy value of the set under each transaction data item, which requires computation using the algorithm in equation (5).

b) The sum of the fuzzy values is calculated.

c) If the result at the current level is not less than the minimum support K, then the set is placed in the frequent itemset.

d) Select all rules that satisfy confidence not less than the predefined confidence threshold T, where T is the predefined minimum confidence.

---

2 Example Analysis of Multi-Level Fuzzy Association Mining

To concretely illustrate the algorithm' s application process and effectiveness, this section presents an empirical analysis using a specific example. The example uses product sales from a fast-moving consumer goods (FMCG) retail supermarket as transactions. To simplify the verification process, seven transactions were randomly selected, as shown in Table 1.

FMCG Sales Transaction Definition Table

Using the predefined taxonomy, their classification is shown in Figure 3.

[Figure 3: see original paper] Predefined classification

As shown in Figure 3, terminal retail stores' fast-moving consumer goods are divided into three categories: personal care products, pet products, and food & beverages. Each category can be further subdivided into several subcategories to identify industry segments and corresponding brands.

For each FMCG category, there is a unique membership function. Based on the membership function calculation results, each item can be further divided into three fuzzy regions: low, medium, and high membership.

First, the FMCG node classification shown in Figure 3 is converted to its equivalent encoding, with results shown in Table 2.

Encoded transaction data for the example

Let the two variables and for this hierarchical structure both equal 1, where represents the classification level of the current item and represents the number of items in the current frequent itemset.

All transactions in the database with similar are merged into a large category and summed. For example, items (223,2) and (254,2) can be integrated into (2**,4). The results of this operation are shown in Table 3.

Level 1 representation in the example

According to the corresponding membership functions, the obtained groups are converted into fuzzy set form. Taking (1**,8) as an example, according to the predefined classification in Figure 3, this group belongs to the personal care products category and requires using the aforementioned personal care products membership function. In this membership function, the calculation result is 6, corresponding to a low region membership of 0.6, medium region membership of 0.9, and high region membership of 0.1. Through this method, equivalent fuzzy sets composed of all items in transactions can be calculated.

The sum of each fuzzy region value is calculated across all transactions to obtain the sum of fuzzy region memberships, as shown in Table 4.

Sum count table for level 1 fuzzy region memberships across transactions

Based on the calculation results summarized in Table 4, the fuzzy region with the highest value in each group is selected. After completing the previous step, the membership of the selected fuzzy region in each group is compared with the predefined minimum support for level and added to the frequent itemset.

For example, assuming the minimum support for level 1 is 1.3, Table 4 shows that 1**.medium, 2**.medium, and 3**.low are all greater than or equal to 1.3. These frequent member sets are placed in the frequent itemset. The candidate itemset is generated from the frequent itemset. Since the frequent itemset consists of three members–1**.medium, 2**.medium, and 3**.low–the members of the candidate itemset are shown in Table 5.

Candidate itemset for level 2

For each 2-member itemset in the candidate itemset, the following steps are executed:

a) The fuzzy membership of each 2-member itemset is calculated based on the predefined membership functions for each item in the transaction. Taking the itemset as an example, this set' s membership in transaction can be calculated according to equation (8).

b) The sum of fuzzy memberships for each 2-member itemset in the candidate itemset can be calculated using the method from part A.

c) Based on the obtained itemsets, only (2**.medium, 3**.low) yields a result greater than the predefined minimum support of 1.3 for level 1. Therefore, the frequent itemset contains only this member. Let s=2, where s represents the number of items in the current itemset. Since there is only one 2-member itemset, a 3-member itemset cannot be generated at level 2. A unit is added to , and step b) is executed.

Let the minimum support for levels 2 and 3 be . The frequent itemsets for these two levels are shown in Table 7 and Table 8, respectively. Since level 4 does not exist, the next step can be executed directly.

Frequent itemsets for level 2
Frequent itemsets for level 3

Based on the frequent itemsets obtained from the previous steps, fuzzy association rule mining can be conducted: retrieve all possible rules from the frequent itemsets at each level according to the following rules. Note that rules must be extracted from frequent itemsets containing at least binary items. The specific rule set is as follows:

- If 2**=medium then 3**=low

- If 3**=low then 2**=medium
- If 3**=low then 21*=medium
- If 21*=medium then 3**=low
- If 211=medium then 3**=low
- If 3**=low then 211=medium

To obtain rules that meet user-specified conditions, the confidence of each rule must be calculated, with results shown in Table 9.

Confidence of all rules

Comparing all rule confidence values in Table 9 with the predefined minimum confidence threshold, rules with confidence greater than the threshold are retained as final mining results. For example, if the minimum confidence threshold is set to 0.9, the final rules are:

- If 2*=medium then 3**=low
- If 3**=low then 2*=medium
- If 211=medium then 3**=low
- If 3**=low then 211=medium

[Figure 4: see original paper] Number of association rules under different minimum confidence values

An important advantage of the proposed multi-level fuzzy association mining algorithm is its ability to mine association rules at different levels according to user needs. In the proposed algorithm, users can precisely specify which level of rules to mine, ensuring the results maximally satisfy user requirements. This is because different levels in the algorithm can have separately defined minimum support values. By increasing the minimum support value at a particular level, the number of rules mined at that level can be reduced to zero. Figures 5 and 6 show the relationship between the number of rules mined at structural levels 1 and 2 and the predefined minimum support across 1,000 transactions. The figures demonstrate that compared to other methods, the proposed algorithm can more accurately obtain rule quantities under different minimum support values—that is, it can mine a narrower range of qualified products with higher precision. The minimum confidence value is set to 0.2.

[Figure 5: see original paper] Number of mined rules corresponding to different minimum support values at level 1
[Figure 6: see original paper] Number of mined rules corresponding to different minimum support values at level 2

3 Experimental Comparison and Analysis

This paper proposes a data mining algorithm based on multi-level fuzzy association rules and elaborates on its application process through specific examples, verifying its feasibility and effectiveness in practical applications. To further analyze the algorithm' s performance superiority under large-scale data conditions, the proposed algorithm was used to analyze 1,000 convenience store shopping receipt records on a PC through MATLAB simulation platform. The PC configuration is i5 Core dual-core with 8 GB RAM.

All products in the convenience store can be divided into six categories, each with a predefined membership function. Based on the information on shopping receipts and predefined data, association rules between these items are mined. The predefined categories have six nodes at level 1, representing product names in the test; 12 nodes at level 2, representing subcategories or other specific product category information; and 45 nodes at level 3, representing manufacturers of these products.

Transaction information on shopping receipts includes product name, model, unit price, and purchase quantity. In each transaction, the same item cannot be included multiple times.

Figure 4 shows the relationship between the number of rules mined and different predefined minimum confidence values across 1,000 transactions with a minimum support of 3. As the number of transactions increases, the number of mined rules also gradually increases. This is because the number of frequent items increases with transaction volume, enabling association mining to discover more rules under a fixed minimum support. Meanwhile, the results indicate that increasing the predefined minimum confidence value reduces the number of mined association rules.

According to surveys on association mining algorithm application requirements, algorithm running time and computational resource consumption are key factors considered by users. If algorithm running time is excessively long, it largely loses its appeal to users. Figure 7 shows a performance comparison of running time between the proposed algorithm and other algorithms when processing different transaction volumes. In the algorithms proposed in literature [5,7], all classification standards are defined by single values, and minimum support and membership functions correspond to all items. To ensure controlled variable comparison, rules at levels 1 and 2 are mined with minimum confidence ranging from 0 to 6. Figure 7 presents the algorithm running time comparison results, showing that the proposed algorithm has shorter running times than the methods in literature [5,7] and [10] under different minimum support values. This is because the proposed method uses improved mining of association rules to more efficiently obtain universal and applicable rules. The proposed algorithm

can not only mine rules at different levels according to user intentions but also reduce program running time, significantly enhancing user satisfaction.

[Figure 7: see original paper] Algorithm running time comparison under different minimum support values

4 Conclusion

Based on a detailed elaboration of the research status and main challenges of existing association rule mining algorithms, this paper proposes a mining algorithm based on multi-level fuzzy association rules by comprehensively utilizing fuzzy set theory, multi-level structural classification methods, and data mining theory. The algorithm can extract implicit information from quantitative data. Using high-frequency itemsets and forming a top-down mining process through progressively deepening iterations, the method can mine association rules at different levels according to user preferences and define different membership functions for different items, thereby meeting the customized classification needs of various product types.

Through association mining experiments on historical databases of FMCG terminal stores and convenience stores, the proposed multi-level fuzzy association rule algorithm demonstrates higher mining precision and significantly reduced computational time compared to related research results, contributing to improved user satisfaction.

References

[1] Gu Yuping, Cheng Longsheng. Research on unbalanced data classification based on MTS-AdaBoost [J]. Application Research of Computers, 2018, 35 (2): 346-348.

[2] Cui Yihui, Song Wei, Wang Zhanbing, et al. A lattice-based clustering data mining method for privacy preservation [J]. Journal of Software, 2017, 28 (9): 2293-2308.

[3] Pemajayantha V, Pemajayantha V, Mellor R, et al. Approaches of discriminant analysis for data mining and management [J]. Science, 2018, 200 (4349): 1481-1483.

[4] Wijayanti A W. Analisis hasil implementasi data mining menggunakan algoritma apriori pada apotek [J]. 2017, 3 (1): 60-69.

[5] Xie Zhiming, Wang Peng. A parallel matrix a priori algorithm based on MapReduce architecture [J]. Application Research of Computers, 2017, 34 (2): 401-404.

[6] An Xianghua, Yu Jingbo, Cai Weiguo. Fuzzy rough FMEA evaluation method based on mixed multi attribute decision making and association analysis

[J]. Computer Integrated Manufacturing System, 2016, 22 (11): 2613-2621.

[7] Zhao Xuejian, Sun Zhixin, Yuan Yuan. An efficient association rule mining algorithm based on pre-selection [J]. Journal of Electronic and Information Science, 2016, 38 (7): 1654-1659.

[8] García S, Luengo J, Herrera F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining [J]. Knowledge-Based Systems, 2016, 23 (7): 98: 1-29.

[9] Ghaffarian S M, Shahriari H R. Software vulnerability analysis and discovery using machine-learning and data-mining techniques: a survey [J]. ACM Computing Surveys, 2017, 50 (4): 1-36.

[10] Yang J, Li J, Liu S. A new algorithm of stock data mining in Internet of multimedia things [J]. Journal of Supercomputing, 2017 (9): 1-16.

[11] Xu Donghao, Li Hongwei, Zhang Tieying, et al. Mining association rules using improved particle swarm optimization [J]. Surveying and Mapping Science, 2016, 41 (2): 168-172.

[12] Zhang X, Ding S, Sun T. Multi-class LSTMSVM based on optimal directed acyclic graph and shuffled frog leaping algorithm [J]. International Journal of Machine Learning & Cybernetics, 2016, 7 (2): 241-251.

[13] Yang, Yang Shulue, Ke Min. Fuzzy sorting search scheme based on Simhash under encrypted cloud data [J]. Acta Computer Science, 2017, 40 (2): 431-444.

[14] Lee J, Kim H, Kim N R, et al. An approach for multi-label classification by directed acyclic graph with label correlation maximization [J]. Information Sciences, 2016, 351 (7): 101-114.

[15] Baxter J S H, Rajchl M, Mcleod A J, et al. Directed acyclic graph continuous max-flow image segmentation for unconstrained label orderings [J]. International Journal of Computer Vision, 2017, 123 (3): 415-434.

Note: Figure translations are in progress. See original paper for figures.