Advances in Intelligent Systems Research, volume 163

8th International Conference on Management, Education and Information (MEICI 2018)

Improved Algorithm of Association Mining and Classification Fusion based on Temporal Interval Lattice

Qing Tan[1, a]

[1]College of Information Technology, Luoyang Normal University, Henan Luoyang, 471934, China

[a]edutanqing@163.com

Keywords: Association mining; Classification; Clustering; Temporal interval; Data mining

Abstract. This paper first describes the combined application of association rules data mining algorithm under temporal constraints. Clustering is to classify a group of individuals into several categories according to similarity. The purpose of clustering is to make the differences between individuals belonging to the same category as small as possible. The paper presents improved algorithm of Association Mining and Classification Fusion based on temporal interval Lattice. The database structure includes at least three fields: transaction number, temporal interval and item sequence. The temporal interval of this paper reflects the time range of the occurrence or collection of corresponding item sequences.

Introduction

The basic framework and process of data mining system has become clear, but affected by application fields, mining data types and knowledge expression patterns, in the specific implementation mechanism, technical route and various stages or components (such as data cleaning, etc.), The functional orientation of knowledge formation and pattern evaluation still needs to be further studied [1]. Because data mining is to discover potential and unknown knowledge in a large number of source data sets, it is inevitable to interactively explore mining with users. This interaction may occur at different stages of data mining, interacting from different angles or at different granularity. So a good interactive mining system is also a prerequisite for the success of the data mining system.

Clustering is the process of grouping the set of physical or abstract objects into multiple classes or clusters, which makes the objects in the same cluster have higher similarity, but the objects in different clusters are different. Clustering is different from classification. Clustering is unknown and can be classified according to known rules. Clustering is an unsupervised learning that does not rely on predefined classes and training examples with class labels. It belongs to observational learning.

In time series, the local extremism points have a relatively important function, but some adjacent local extremism points have smaller time intervals and no difference in numerical values. These local extremism points do not play a key role in the compression of time series. At the same time, people will not pay much attention to these extreme points, so they are called the local extremism points as non. Important local is extremism points. In all the local extreme points of time series, the key points are removed from the unimportant local extreme points [2].

The task of discovering association rules is to find the strong rules with confidence degree, support degree equal to the given value from the database, data classification and statistics, rough set and so on. Linear regression and linear discriminate analysis are typical statistical models. In order to reduce the cost of decision tree generation, an interval classifier is proposed. Recently, neural networks have been used to classify and extract rules in databases, which is represented by backward propagation of classification.

The case based reasoning classification is based on the requirements. Unlike the nearest taxonomy, the training sample is stored as a point in the Euclidean space. The sample or "case" stored in CBR is a complex symbol description. It tries to combine the adjacent training cases and put forward the solution of the new case. Case based reasoning may use background knowledge and problem solving strategy.

The challenge of case based reasoning includes finding a good similarity measure and developing an effective technique and a combined solution to the training case index.

The types of data objects processed in cluster analysis generally include interval scale variable, symmetric binary variable, asymmetric binary variable, nominal variable, ordinal variable, proportional scale variable and mixed variable.

Analysis of Association Rule data Mining Algorithm under Temporal Constraints

ID3 selects attributes using the information gain of the subtree. Here we can define the information by using entropy and entropy, which is a measure of impurity, that is, the change of entropy. C4.5 uses the information gain rate [3]. Yes, the difference is that one is the information gain; the other is the information gain rate.

Neural network, as a relatively independent branch of research, has been put forward very early, and its principle has been introduced in detail in many works and literatures. It is difficult to apply neural network because of its long training time and its poor interpretability. However, the neural network has the advantages of high anti-interference ability and the ability to classify untrained data, which makes it attractive. Therefore, the use of neural networks in data mining is a meaningful but still hard work to explore.

The root node of the decision tree is the attribute with the largest amount of information in all samples. The middle node of the tree is the attribute with the largest amount of information in the sample subset contained in the subtree with the root of that node. The leaf node of the decision tree is the class value of the sample. (how to classify) decision trees are used to classify new samples, that is, by testing the values of new sample attributes in the decision tree, starting from the root node of the tree, according to the values of the sample attributes, the decision tree is gradually down the decision tree to the leaf node of the tree. The category represented by the leaf node is the category of the new sample. Decision tree method is a very effective classification method in data mining.

Because of the real data is often with missing, inconsistent and noise. Therefore, in order to improve the efficiency of data mining process and the quality of results, data preprocessing is needed, as is shown by equation(1), u is time series preprocessing is mainly about cleaning noise data in time series [4].

$$u(k) = k_p(e(k) + \frac{T}{T_i}\sum_{j=0}^{k} e(j) + \frac{T_d}{T}(e(k) - e(k-1))) \tag{1}$$

Most of the existing association rules mining algorithms use the framework of support confidence threshold. Although such a threshold framework can eliminate a large number of boring rules, there are still some that exist, and the correlation metric is used to extend the framework. In order to ensure the accuracy of association rules mining results, the importance threshold is introduced to further shield the boring rules. Define importance.

Data association is a kind of important knowledge that can be found in database. If there is some regularity between the values of two or more variables, it is called correlation. Correlation can be divided into simple correlation, temporal correlation, and causal correlation. The purpose of association analysis is to find the hidden association network in the database. Sometimes the association function of the data in the database is not known, even if it is known; the rules generated by the association analysis have confidence and support.

Aiming at the shortcomings of FP-Growth algorithm, this paper improves the classical algorithm, and proposes a method of using support counting two-dimensional table, thus eliminating the first traversal of conditional schema base by classical algorithm. The algorithm is described as follows: At the same time, a two-dimensional vector is created at the same time that the transaction set T is traversed for the first time, and the count of the support degree of each item combination in each transaction is recorded. If there is a business "A / B / C / D", then in the two-dimensional vector table, it is necessary to add to the item "A / B" / (A / C / D). Among them, Vectors C (C) and B (C) are two

different vectors. 2 when creating conditional FP subtrees under (conditions≠ {Null}) by recursion, there is no need to traverse the conditional schema bases twice (where the first traversal of conditional schema bases can obtain a list of support counts.

In view of these two incomplete data characteristics, the following two ways are adopted to fill the data in data cleaning: the missing attribute value is replaced by the same constant, such as "unknown". This method is used to deal with the data of the first data feature mentioned above by replacing the null value with a replacement value. The processed data will be deleted without any value to the later mining work. Fills the missing has value with the most likely value of this property. For the data of the second kind of data feature, the value of each attribute is counted in advance, the distribution state and frequency of the value are counted, and all the missing values of the attribute are filled with the value with the highest frequency.

Improved Algorithm of Association Mining and Classification Fusion based on Temporal Interval Lattice

Traditional clustering methods can only identify convex and spherical clusters, but in fact, many clusters are concave and have complex shapes, so accurate identification of complex clusters is still one of the hot topics. In addition, as the size of the database increases, the cluster in the database is no longer a single feature, that is, there are clusters of different sizes, different densities, different shapes in the database, and sometimes the difference between clusters and clusters is not obvious. How to identify these clusters is also to be studied in [5]. Almost all clustering studies are conducted on the basis of validity, because this is the purpose of clustering.

Later stage: since time is at this point > maxgaps, it is necessary to search for xj-1 from the time value of time (xj)-margay, but at the same time keep the xj-2 position unchanged. When the newly found xj-1 is still not satisfied with the maxgap, the xj-2 is searched again after the time value is time (xj-1)-maxgap, and the xj-3 position remains the same until the location element xj-i satisfies the condition or the x1 cannot keep the same position [6]. At this time, the forward phase is returned.

The k-means algorithm is a clustering algorithm, which divides n objects into k partitioned k < ngs according to their attributes. It is similar to the maximum expectation algorithm dealing with the mixed normal distribution (the fifth of the top ten algorithms) because they are trying to find the center of the natural clustering in the data. The object attributes are assumed to come from space vectors and the goal is to minimize the sum of mean square errors within each group.

Clustering is to classify a group of individuals into several categories according to similarity. The purpose of clustering is to make the differences between individuals belonging to the same category as small as possible, while the differences among individuals in different categories are as large as possible. One of the goals of data mining is clustering analysis. Through clustering technology, records in the source database can be divided into a series of meaningful subsets, and then the data can be analyzed. For example, as is shown by equation (2), x is a commercial sales company may care which customers are more interested in a given promotion strategy [7].

$$\begin{cases} \left|\delta_{(k+1)k} - \delta_{k(k-1)}\right| > \varepsilon \\ \left|x_k - x_{k-1}\right| > \delta \\ \left|x_k - x_{k+1}\right| > \delta \end{cases} \tag{2}$$

K- center point algorithm: (input) the number of results clusters K, including the data sets of n objects. (output) K clusters, which make the sum of the difference of all objects and their nearest centers minimum. (cluster process) (1) randomly select k object as the initial center point; (2) calculate the distance between other objects and the k center, and then return each object into the distance. Its "nearest" cluster; (3) randomly select a non central point object Orandom, and calculate the total cost S for Orandom instead of Oj; (4) if S<0, then use Orandom instead of Oj to form a new set of k center points; (5) repeat the iteration 3,4 step until the center point is not constant.

The measurement of pattern similarity in time series is the basis of obtaining frequent patterns in pattern sequences. Only by measuring the similarity between patterns can the frequent pattern acquisition and the generation of strong temporal association rules in pattern sequences be better completed.

Records in the database can be divided into a series of meaningful subsets, that is, clustering. Clustering enhances people's understanding of objective reality and is a prerequisite for conceptual description and deviation analysis. Clustering technology mainly includes partitioning method, hierarchical method, density-based method and model-based method. Some clustering algorithms inherit the idea of many clustering methods.

Data integration is the integration of data from different sources, formats and characteristics, which is logically or physically integrated to provide a complete data source for data mining.

System Experiments and Analysis

The DENCLUE method uses an influence function to simulate the effect of each data point in the domain, and the sum of the influence functions of all the data points to simulate the overall density of the data space [8]. The main advantage of the cluster .DENCLUE method is that it can deal with high-dimensional data, concentrate arbitrary clusters, and has strong anti-noise ability and faster processing speed by determining the local maximum of the density attraction point, I. e., the global density function. Its disadvantage is that the input density parameter σ and the noise threshold ε are required, and the clustering results are sensitive to these two parameters.

In the two stages of examining whether a data sequence d contains a candidate k sequence s, it is necessary to continuously search for a single element in the candidate sequence s in the data sequence d. Therefore, the data sequence d is transformed as follows: for each item in d, a list of the time of occurrence is established [9]. In this case, if we want to find the transaction time corresponding to the first occurrence of an item x after the transaction time t, we only need to traverse the transaction time list of x until we find a transaction time greater than t. If we want to find out the first occurrence of an element of candidate sequence s (x1xn) after transaction time t, we only need to traverse the transaction time list of each item xi($1 \leqslant i \leqslant n$) to find out the transaction time that Xi first appears after transaction time t.

FCM takes n vectors is divided into c fuzzy groups, and the clustering centers of each group are obtained. The main difference between FCM and HCM is that the value function of dissimilarity index reaches the minimum. The main difference between FCM and HCM lies in the fuzzy partition of FCM. The degree of membership of each given data point to each group is determined by membership between 0 and 1. In accordance with the introduction of fuzzy partitioning, as is shown by equation (3), G is the membership matrix U allows elements with values between 0 and 1.

$$G(s) = \frac{U(s)}{E(s)} = k_p\left(1 + \frac{1}{T_1 s} + T_D s\right) \tag{3}$$

DBSCAN is a representative density-based method, which controls the growth of clusters according to a density threshold. OPTICS is another density-based method, which calculates a clustering order for automatic and interactive clustering analysis.

Given the number of partitions to be constructed, the partition method first creates an initial partition. Then an iterative repositioning technique is used to improve the partition by moving objects between partitions. A general criterion for good partitioning is that objects in the same class are as "close" or relevant as possible, while objects in different classes are "far from" or "different" as much as possible, as is shown by equation(4).

$$D(M_i, M_j) = \beta \cdot \left|a_i - a_j\right| + (1-\beta)\left|\Delta t_i - \Delta t_j\right| \tag{4}$$

Through the hierarchical decomposition of the data in the source database, the target cluster can be generated step by step. There are two basic methods: agglomeration and division. Condensed clustering refers to the gradual merging from small to large (starting with each tuple as a group) until each cluster meets the characteristic condition. Splitting clustering refers to the gradual splitting from large to small (possibly a group at first) until each cluster meets the characteristic condition.

Given a database of n objects or tuples, a partitioning method constructs k partitions of data, each representing a cluster and k < n. (hierarchical method) a hierarchical decomposition of a given set of data objects. (density-based method) continue clustering as long as the density of the adjacent region exceeds a threshold. A grid-based approach that quantifies the object space is into a finite number of elements. (model-based approach) attempts to optimize the adaptability between given data and some mathematical models.

The time series is divided into a series of subsequences with key points, and the linear regression of each subsequence is carried out. Finally, the fitting line is obtained. For example, the fitting equation of the sequence is obtained by the least square method.

Summary

In the real world, most of the information that can be obtained is stored in a text database, consisting of a large number of documents from various data sources. The text database has been developed rapidly because of the rapid growth of information in the electronic form. The most stored data in the document database is the so-called semi structured data (semi structure data). It is neither completely unstructured nor completely structured. In recent research on the database domain, a large number of studies on Modeling and implementation of semi structured data have been made.

References

[1] Dhar V, Tuzhilin A. Abstract-driven pattern discovery in databases. IEEE Trans. Knowledge and Data Eng.,2013: 926-938.

[2] Yan Xuesong, Cai Zhihua, an efficient Association Rule Mining algorithm based on Apriori, computer Engineering and Application,2012,10:209-211.

[3] Cheeseman P et al, Bayesian classification (AutoClass):Theory and results. advances in knowledge discovery and data mining, AAAI/MIT Press, 2016: 153-180.

[4] Mobasher B, Cooley R, Srivastava J,Automatic personalization based on web usage mining, Communications of the ACM, 2010，43(8):142-151.

[5] Ming Fan, Niu Chang Yong, Zhu Yan. An effective algorithm for multidimensional association mining. Computer science, 2011,28 (11):44-47.

[6] Srivastava J, Cooley R, Deshpande M, et al. Web usage mining:discovery and application of usage patterns from web data, SIGKDD Explorations, 2010,1(2):12-23.

[7] Peng Zhen, Pei Lili, Yang Bingru. A new method for mining association rules. Computer engineering and applications, 2014, (45):127-129.

[8] Agrawal R et al. The QUEST data mining system. In Proc. Int. Conf. Data Mining and Knowledge Discovery(KDD'96), 2016: 244-249.

[9] Zheng Quan Chao, Wu Jianhua. The algorithm of solving maximal frequent itemsets based on binary search is modified, Computer applications and Software,2010,(5): 269-271.