Data Mining Algorithm to Optimize the Development of Computer Artificial Intelligence Industry

Zhang Limin
Anyang Preschool Educationg College, Anyang, Henan, China

Abstract

In today's big data era, artificial intelligence plays an irreplaceable role in computer network technology. Data mining algorithm is widely used in intelligent industry. This paper studies the process of data mining algorithm to optimize the development of computer artificial intelligence industry. In this paper, the application of artificial intelligence in data mining and data protection, problem solving, microcomputer services, computer network security management are discussed. The main purpose of this paper is to give full play to the advantages of artificial intelligence and provide guarantee for the development of China's computer industry. Therefore, this paper will focus on the corresponding content of artificial intelligence in computer network technology in the era of big data.

Keywords*: Big data, artificial intelligence, network technology, network security management.*

I. Introduction

People regard raw data as a source of knowledge, just like mining from ore. The original data can be structured, such as data in relational database, or semi-structured, such as text, graphics, image data, or even heterogeneous data distributed on the network [1-2]. The method of discovering knowledge can be mathematical or non mathematical; It can be deductive or inductive. The discovered knowledge can be used for information management, query optimization, decision support, process control, etc., and can also be used for data maintenance. Therefore, data mining is a broad interdisciplinary, it brings together researchers in different fields. Especially the scholars and engineers in database, artificial intelligence, information processing, mathematical statistics, visualization and parallel computing [3]. At the same time, the idea of data mining also points out a new research direction for the development of AI, which makes intelligent information processing have new technologies and means [4-6].

The query, statistics and analysis in the existing technology are verified from a large number of data on the basis of existing assumptions. Data mining is to mine information and discover knowledge (KDD) without explicit assumptions. The results have three characteristics: unknown in advance, effective and practical. Data mining is not to replace the traditional query and analysis technology, on the contrary, it is the extension and expansion of query, statistics and analysis methodology. It is to get new patterns, conclusions and hypotheses (discover the relationship and rules between data) from a large number of data. Data mining is to discover information or knowledge that cannot be discovered by intuition. The more unexpected the information is, the more valuable it is.
The core part of data mining is the process of building models for data sets. Different data mining methods have different ways of building data models. Many different methods can be used in data mining, such as association rule discovery, neural network, decision tree, genetic algorithm and visualization technology, and there are hundreds of derived methods under the same method.

II.Association Rules Mining

2.1 Basic concepts of association rules
The results of association rules mining usually tell people something like this: "among the customers who buy

bread and butter, 90% also buy milk" (buy bread, butter → buy milk).

Association rules were first used to analyze the shopping affairs of supermarket customers. It can help how to put the goods on the shelf (such as putting the goods that customers often buy together) and how to plan the market (how to match each other) [7]. Therefore, the main object used for association rule discovery is transactional database. A transaction is generally composed of the following parts: transaction processing time, a group of elements, and transaction identification number.

2.2 Classification of association rule methods

The existing algorithms of association rules can be roughly divided into search algorithm, hierarchical algorithm, data set partition algorithm, sampling algorithm and so on.

1. Search algorithm: when reading each transaction in the data set, the search algorithm processes all itemsets contained in the transaction, so the search algorithm needs to calculate the support number of all itemsets in the data set D. AIS algorithm, STEM algorithm, and association rule mining algorithm based on grid building algorithm are all such search algorithms.

2. Hierarchical algorithm: The hierarchical algorithm represented by Apriori algorithm is to find frequent itemsets in the order of increasing the number of items. Apriori's algorithm finds out all the frequent k itemsets when scanning the dataset for the k th time. The candidate itemsets in the (k+1) th scan of data sets are generated by all frequent k itemsets through join operations.

3. Data set partition algorithm: data set partition algorithm includes partition algorithm, DIC algorithm, etc. these algorithms divide the whole data set into data blocks that can be stored in memory for processing, so as to save the I / O cost of accessing external memory. The number of candidate itemsets of data set partition algorithm is generally larger than that of Apriori algorithm. Increasing the data distortion of each data block can reduce the number of candidate itemsets. Data set partition algorithm is the basis of various parallel and distributed association rule mining algorithms.

4. Sampling algorithm: the sampling algorithm generates the sampling data set D 'by sampling the data set D, finds out the frequent item set in the sampling data set D' as the candidate item set, and then scans the data set D to determine the frequent item set. How to calculate the negative boundary to retrieve the missing frequent itemsets is the key of the sampling algorithm. Sampling algorithm is suitable for mining association rules in the environment of high efficiency and low accuracy.

2.3 Apriori algorithm

Agrawal put forward an algorithm for mining association rules between itemsets in customer transaction database in 1994-Apriori, which belongs to one-dimensional, single-layer and Boolean association rules in classification [8]. Up to now, Apnon algorithm is still the most influential method for mining frequent itemsets of Boolean association rules, where all itemsets whose support degree is greater than the minimum support degree are called frequent itemsets. The core of Apriori algorithm is based on the recursive idea of two-stage frequent itemsets, using an iterative method called layer-by-layer search. k- itemsets are used to explore (k+l)- itemsets. First, find out the frequent 1- itemsets. This set is denoted as L_1. L_1 is used to find the set L_2 of frequent 2- itemsets, while L_2 is used to find L_3, and so on until the frequent k- itemsets cannot be found. Finding each Lk requires a database scan [9-10].

In order to improve the efficiency of the algorithm, marmila introduced pruning technology to reduce the size of the candidate set C_k and compress the search space. The pruning strategy introduced in the algorithm is based on such a property: all non empty subsets of frequent itemsets must also be frequent. According to the definition, if itemset I does not meet the minimum support threshold, min_ Sup, then I is not frequent, that is P (I) < min_ sup。

If item a is added to I, the result itemset (I ∪ A) cannot appear more frequently than I. Therefore, IUA is not frequent, that is, P (I ∪ A) < min_ sup。 This property belongs to a special classification, called anti monotone, which means that if a set cannot pass the test, all its supersets cannot pass the same test.

Once frequent itemsets are found by transactions in database d, it is straightforward to generate strong association rules from them (strong association rules satisfy minimum support and minimum confidence). For the confidence level, the following formula can be used, where the conditional probability is expressed by the item set support count:

$$confidence\left(A \Rightarrow B\right) = P\left(A\middle|B\right) = \frac{support\_count\left(A \cup B\right)}{support\_count\left(A\right)} \quad (1)$$

In which, support_count(A ∪ B) is the number of transactions including itemset aub, and support_count(A) is the number of transactions including itemset a .. According to this formula, the following association rules can be generated:
(1) for each frequent itemset 3, all non-empty subsets of 1 are generated.
(2) for each non-empty subset s of 1, if $\frac{support\_count\left(1\right)}{support\_count\left(s\right)} \geq min\_conf$,
The output rule is "s = > (1-s)". Where, min_ conf minimum confidence threshold.
Since rules are generated from frequent itemsets, each rule automatically satisfies the minimum support.

2.4 The shortcomings of Apriori algorithm and its improvement
Apriori algorithm needs to repeatedly scan the database, needs a lot of I / O load, and may produce a large number of candidate sets, which are two major shortcomings of Apriori algorithm. A large number of candidate sets may be generated. When there are 10000 frequency sets of length 1, the number of candidate sets of length 2 will exceed 10m. Through experiments, we can find that the main calculation of finding frequent sets is to generate frequent 2-itemsets, and the process of generating frequent 2-itemsets is the bottleneck of Apriori algorithm mining. In addition, if you want to generate a very long rule, you need to generate a large number of intermediate elements.

Although Apriori algorithm itself has been optimized, but in practical application, there are still unsatisfactory places, so people have put forward some optimization methods.

1. Hash based method
An efficient hash based algorithm for generating frequency sets is proposed by Park et al. Through experiments, we can find that the main calculation of finding frequency set is to generate frequent 2-itemsets L_k. Park et al. Use this property to introduce hash technology to improve the method of generating frequent 2-itemsets.

2. Dynamic itemset count
Dynamic itemset counting is to add candidate itemsets at different times of scanning. Dynamic itemset counting is proposed when mining database partition. Each data block divided is marked with a start mark. In this change, a new candidate set can be added at any starting point: the candidate set has been determined before each database scan. This technique is dynamic because it estimates the support of all itemsets counted so far; If all subsets of an item set are estimated to be frequent, a new candidate item set is added. So the algorithm needs to scan twice.

3. Sequence mode
Sequence pattern is the discovery of transaction sequence (pattern) that changes with time. The purpose of sequence pattern analysis is to mine the sequence of occurrence of item sets which are mainly in the sequence of occurrence time. This kind of research is less in practice.

4. Analysis of cyclical market shopping
The periodic market shopping analysis is to find the corresponding frequent item set in the user-defined cycle. The

cyclical market shopping analysis uses time-marked transaction records to determine the subset of the transaction database and mark it as cyclical. The so-called cycle is a set of such as "the first day of each month, etc. the corresponding association rules are extracted from the periodic daily items. Therefore, a set of items that do not meet the minimum support for sniffle may be considered frequent in the data subset that satisfies periodic constraints.

5. Other methods include mining multi-level multi-dimensional association rules and mining of time series data, etc.

III. Research on Clustering

3.1 The basic concept of clustering

Clustering is a process of dividing data sets into several groups or classes, and makes the data objects in the same group have high similarity. The data objects in different groups are not similar. Similar or dissimilar description is determined based on the value of data description attribute. It is usually represented by the distance (between objects). Clustering methods include statistical method, machine learning method, neural network method and database oriented method.

In statistical methods, cluster analysis is called cluster analysis, which is one of the three methods of multivariate data analysis (the other two are regression analysis and discriminant analysis). It mainly studies clustering based on geometric distance, such as Euclidean distance, Minkowski distance and so on. The traditional statistical cluster analysis methods include systematic clustering, decomposition, addition, dynamic clustering, ordered sample clustering, overlapping clustering and fuzzy clustering. This clustering method is based on global comparison, it needs to examine all individuals to determine the classification. Therefore, it requires that all data must be given in advance, instead of adding new data objects dynamically. Clustering analysis method does not have linear computational complexity, so it is difficult to apply to the case of very large database.

3.2 Types of clustering

Clustering is also a more research direction in data mining, there are many methods, mainly in the following categories.

1. The method of division

Typical division method: k-average, k-center.

(1) K-average method: k-average algorithm takes k as the parameter, divides n objects into k clusters, the objects in the cluster have high similarity, and the similarity is calculated according to the average value of the objects in a cluster.

(2) K-center method: k-center method selects the most central point in the cluster as the reference point. The clustering strategy of k-center method is as follows: firstly, randomly select a representative object for each cluster; The remaining objects are assigned to a cluster according to their distance from the representative objects; Then, the representative objects are replaced by non representative objects repeatedly to improve the quality of clustering.

2. Hierarchical approach

There are two kinds of hierarchical methods: Condensed hierarchical classification and split hierarchical classification.

(1) Hierarchical classification of aggregation: from bottom to top, each object is regarded as a cluster, and then gradually merged into larger and larger clusters until a termination condition is met.

(2) Split hierarchical classification: this method takes all objects as a cluster from top to bottom, and then gradually splits them into smaller and smaller clusters until a termination condition is satisfied.

The difficulty of hierarchical classification method is that the split or merged points are difficult to determine, and the split or merged clusters can not exchange objects or fall back, so the scalability is not good. There are several improved methods, such as birch method. Firstly, the tree structure is used to divide the objects into different levels,

and then other clustering algorithms are used to refine the clustering results. There is also cure, which uses a fixed number of representative objects to represent each cluster, and then shrinks them toward the cluster center according to a defined score.

3. Grid based clustering method

The grid based clustering method uses a multi-resolution grid data structure. It transforms the space vector into a finite number of elements, which form a grid structure, and all clustering operations are carried out on the grid. The advantage of this kind of method is that the processing speed is fast, and the processing time is independent of the number of data objects, and only depends on the number of cells in each dimension of the quantization space.

(1) Sting (statistical information grid) statistical information grid: the spatial region is divided into rectangular cells by sting method. Different levels of resolution correspond to different levels of rectangular cells. These cells form a hierarchical structure: each cell in the high level is divided into multiple cells in the low level. The properties of each grid cell are calculated in advance. For an object, it first determines which rectangular cell it belongs to at the highest level, and then finds the cell of the object in the sub cell of the cell, until it finds which cell the object belongs to at the lowest level.

(2) WaveCluster: the wavelet transform clustering method first aggregates data by imposing a multidimensional grid structure on the data space, and then uses a wavelet transform to transform the original feature space to find dense areas in the transformed space.

IV.Test and Application

Construct and improve the data mining algorithm of association rules, classification prediction and clustering, and design the program: establish a database, apply various algorithms to it, and analyze the results. System software and hardware environment:

(1) Hardware environment: Pentium IV 2.8GHZ, 512MB memory.

(2) Software environment: WindowsXP,JDKl.4.2. iDA(iDataAnalyser),Weka.

4.1 Data mining of students' scores in National Computer Rank Examination

The information of students in school exists in two database files, one is the National Computer Rank Examination score database. Name ID number, name, gender, ID card number, contact number, written examination results, boarding test scores and total score. The other is the natural condition database of students, including student ID number, name, professional class, ID card number, native place and so on. Through the extraction of data from these two databases, we can find out which majors have higher passing rate in which subjects.

After preprocessing the data, the implementation of k-cluster algorithm based on clustering, data mining, found that the distribution of candidates in various examination subjects as shown in Figure 1, the distribution of professional and examination subjects, results as shown in Figure 2. The distribution of examination subjects, majors and scores is shown in Figure 3. From Figure 1 and Figure 2, we can see that the passing rate of Chinese, secretary, sports, Korean, Japanese, Russian, painting, art, dance and other majors applying for Level 2 access is relatively high.

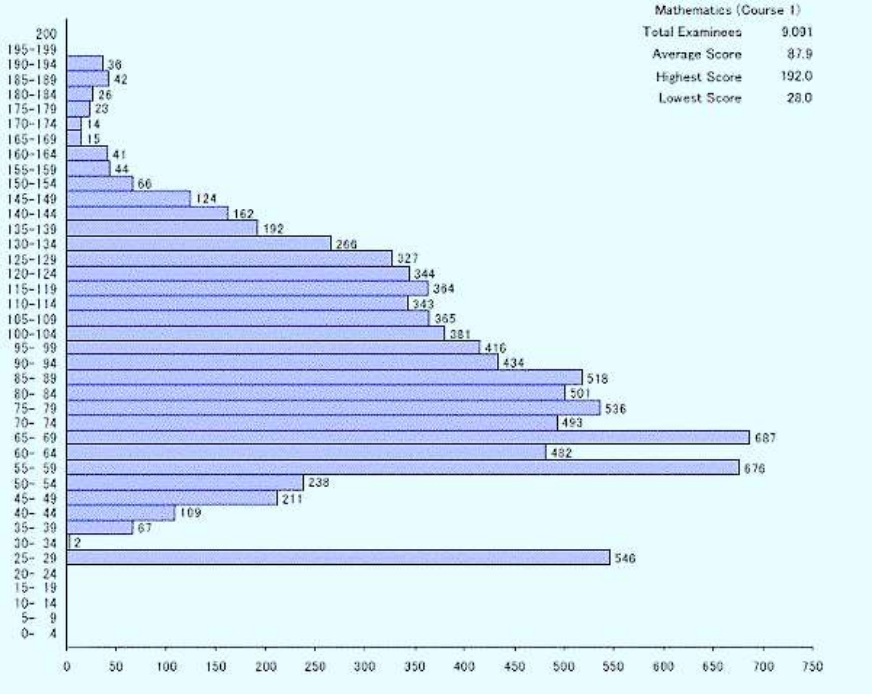

Fig 1:Distribution of examinees in each examination subject

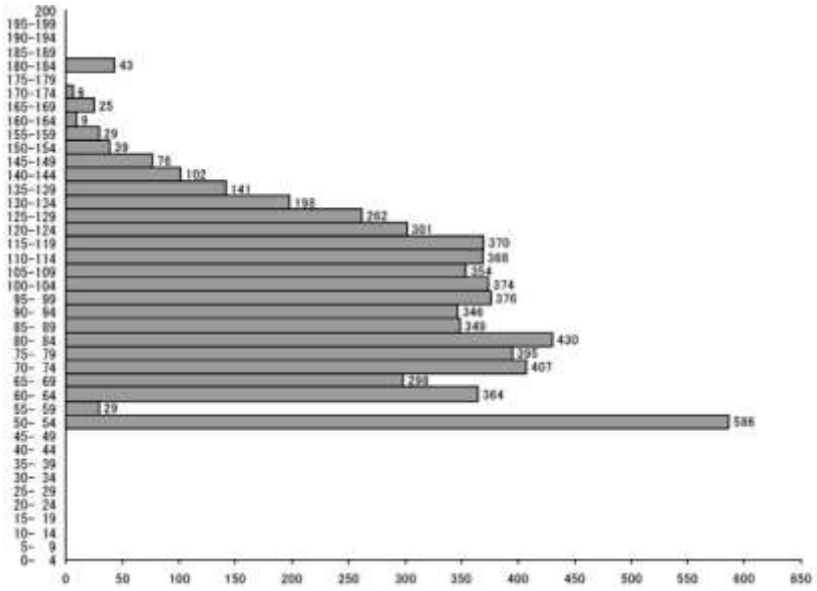

Fig 2:Distribution map of major and examination subjects and scores

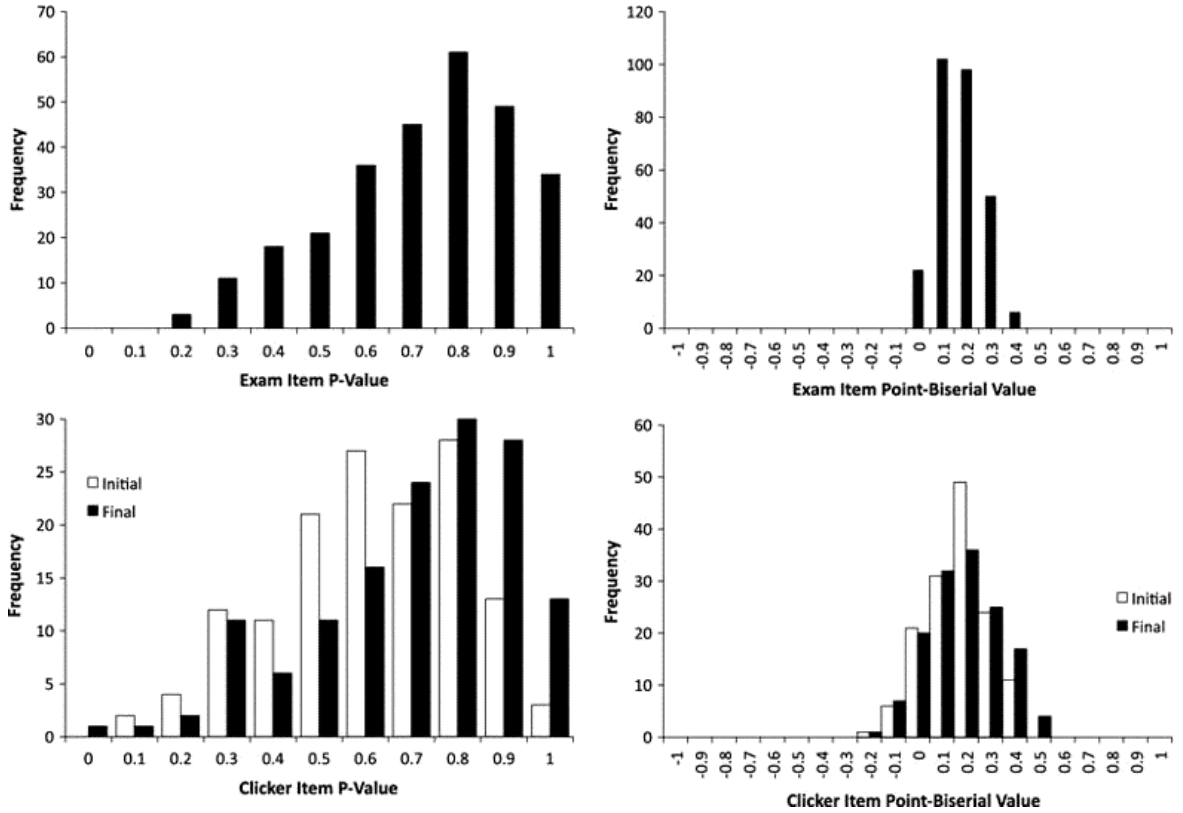

Fig 3: Distribution of examination subjects, majors and grades

4.2 Large supermarket data mining

In the sales records of large supermarkets, each record contains a list of goods purchased once, and the association in the association algorithm can tell us the relationship between two or more goods. For example, 80% of customers buy bread and milk, and 60% of them buy milk as well as bread. We express the relationship between bread and milk as follows: bread → milk (60%, 80%). The association of data items can also be generated among multiple items, such as bread, milk → sauce (60%, 40%), etc.

Because the data of large supermarket is not easy to obtain, a group of simulated data of supermarket sales is used for mining. Data includes transaction number, membership card number, commodity name, unit price, quantity, total amount, time, etc. The same transaction number indicates that it is a transaction activity of a customer, and the purpose of mining is to find whether there is a certain association in the goods purchased by the customer.

1. Data preprocessing

Because there is no defect data in the simulation data of large supermarkets, it is not necessary to make up for them. However, for the transactions with the same sales date and the same membership card number, they should be combined. As a customer's transaction activities on the same day, after sorting out the original data, the records of whether the five commodities (S1, S2, S3, S4, S5) are purchased simultaneously are generated.

2. Find association rules

The Apriori parallel algorithm based on association rules is more suitable for the analysis of customer transactions in large supermarkets. After the implementation of the algorithm to mine the simulated supermarket data, it is found that the rules are as follows, which are divided into five categories.

3. Analyze association rules

The mining association rules show that the probability of buying S3 and S4 but not S1 is 80.57%, and the probability of buying S1 but not S4 is 76.63%. Therefore, supermarkets can put S3 and S4 together to promote sales.

V. Conclusion

This paper introduces the related theoretical knowledge of data mining, including the concept of data mining, classification of data mining, knowledge pattern of data mining and steps of data mining. This paper discusses Apriori algorithm, its improved correlation algorithm and parallel association rule mining algorithm, and summarizes classification method and clustering method. On this basis, the core technology of data mining is successfully applied in intelligent information processing system. In the data mining stage, Apriori parallel algorithm based on association rules, classification based decision tree induction algorithm and clustering based k-cluster algorithm are improved respectively. Through the system operation and test results, the improved algorithm is successful.

References

[1] ELGAFY, ANWAR M.: Environmental Impact Assessment of Transportation Projects: An Analysis Using an Integrated GIS, Remote Sensing, and Spatial Modeling Approach. Environmental Modelling & Software, 2005, 79(C):85-95.

[2] VIRTANEN T , MIKKOLA K , NIKULA A.: Satellite image based vegetation classification of a large area using limited ground reference data: A case study in the Usa Basin, north-east European Russia. Polar Research, 2006, 23(1):51-66.

[3] YANG X , ZHENG Y , GENG G.: Development of PM 2.5, and NO2, models in a LUR framework incorporating satellite remote sensing and air quality model data in Pearl River Delta region, China. Environmental Pollution, 2017, 226:143-153.

[4] FA-WANG Y E , DE-CHANG L.: Application of High Resolution Satellite Remote Sensing Technology in Identification and Analysis of the Uranium Mineralization Bleached Alteration. Remote Sensing for Land & Resources, 2012, 24(4):232-232.

[5] VADREVU K P , LASKO K , GIGLIO L.: Analysis of Southeast Asian pollution episode during June 2013 using satellite remote sensing datasets. Environmental Pollution, 2014, 195:245-256.

[6] ZORAN M , ZORAN L F , DIDA A.: Satellite remote sensing image based analysis of effects due to urbanization on climate and health. Proceedings of SPIE - The International Society for Optical Engineering, 2013, 8893(6):909-927.

[7] FERRIER G.: Application of Imaging Spectrometer Data in Identifying Environmental Pollution Caused by Mining at Rodaquilar, Spain. Remote Sensing of Environment, 1999, 68(2):125-137.

[8] LEIFER I , MELTON C , TRATT D M.: Remote sensing and in situ measurements of methane and ammonia emissions from a megacity dairy complex: Chino, CA. Environmental Pollution, 2017, 221:37-51.

[9] WU X , LIU T , CHENG Y.: Dynamic monitoring of straw burned area using multi-source satellite remote sensing data. Transactions of the Chinese Society of Agricultural Engineering, 2017, 33(8):153-159.

[10] HUANG Y , ORGAN B , ZHOU J L.: Emission measurement of diesel vehicles in Hong Kong through on-road remote sensing: Performance review and identification of high-emitters. Environmental Pollution, 2018, 237:133-142.