Review of Algorithms for Data Mining

Huanchen Bai[1, a] and Xiaojun Liu[1, b*]

[1]School of Electronic & Information, Huanggang Normal University, Hubei Huanggang, China

[a]18623582@qq.com, [b]whutliuxiaojun@126.com

*The corresponding author

Keywords: Data Mining; Classification; Clustering; Association rules

Abstract. Data mining also known as knowledge discovery in databases. This Paper is divided main data mining method into three types: classification, clustering and association rule mining, and gives each method typical algorithm to analyze the advantages and disadvantages of its application scenarios.

Introduction

Technically, Data mining is a lot of, incomplete, noisy, fuzzy, random data is extracted from the implicit therein, it is not known in advance, but is potentially useful information and knowledge the process of. Data mining has become an interdisciplinary, it is mainly the use of statistics, artificial intelligence, machine learning , database technology and other methods to find the model and structure of the data and found valuable relationships of knowledge . Currently, there are the primary method of data mining classification, clustering and association rule mining and other types. [1].

Classification

Classification is an important technology in data mining applications is extremely broad, it has been proposed many algorithms. Classification also known as Supervised Learning. Definition of supervised learning is: given a data set D, supervised learning goal is to produce a set of attribute values contact and Class A standard (a class attribute value called a class standard) set C classification / prediction function that can used to predict the new set of properties (instance data) class standard. This function is called a Classification Model, or Classifier. The main classification algorithm is: decision tree algorithms, rule-based reasoning, Naive Bayes model, support vector machine. [2]

Decision Tree algorithm is the core of Divide-and-Conquer strategy, which uses a top-down recursively constructed tree. There are several ways to choose segmentation, but the goal is same: try to target class optimal segmentation. From root to leaf node has a path, this path is a "rule". A decision tree can be converted into a rule set, the rule set for classification. C4. 5, is a classic decision tree classification algorithm, which is the core of the decision tree algorithm ID3 algorithm improvements. C4.5 algorithm has the following advantages: easy to understand classification rules generated higher accuracy and disadvantages are: the structure of the tree in the process, the need for multiple data sets of sequential scanning and sorting algorithm resulting in inefficiency. In addition, C4.5 only suitable to reside in memory data sets when training set too big to accommodate the program cannot run in memory when.

Rule-based reasoning algorithm is a direct result of the rule set of core rules inference algorithm is Separate-and-Conquer strategy, it evaluates all the attributes - value (condition), then select one. Thus, in one step, Divide-and-Conquer Strategy generates m rule, and Separate-and-Conquer Strategy produced only one rule, much more efficient than the tree, but the basic ideas, the two are identical.

Naive Bayesian Model originated in classical mathematical theory has a solid mathematical foundation and stable classification efficiency. The basic idea is: classification task can be seen to it posterior probability given after a test sample d estimate that Pr (C = cj | d), then we consider the probability of which class cj corresponding to the maximum, they put the category assigned to the

sample d. Naive Bayes classifier probability construct the required value can be obtained through the scan data, so the algorithm relative number of training samples is linear, high efficiency. Meanwhile, NBC needed to estimate model parameters little less sensitive to missing data, the algorithm is relatively simple. Theoretically, NBC model compared with other classification with minimal error rate. But in fact it is not always the case, because the NBC model assumptions are independent properties, this assumption is in practice often is not established, which gives the correct classification NBC model has brought a certain extent. In the large correlation between the number of attributes or more attributes, classification efficiency NBC model not as a decision tree model. And when a property less relevant, NBC performance model is most favorable.

Support Vector Machine, is a supervised learning method, which is widely used in statistical classification and regression analysis. Support vector machine will be mapped to a higher dimensional vector space, the establishment of a maximum interval hyperplane in this space. In separate data on both sides of the hyperplane has two mutually parallel hyperplanes, the two parallel hyperplanes separating hyperplane maximize the distance. Support vector machines not only have a solid theoretical foundation, but in many applications are more accurate than other methods, especially when dealing with high dimensional data. It is by far the most accurate text classification algorithm to solve the problem; it is also widely used in web page classification and bioinformatics.

Clustering

Clustering [3] Also known as Unsupervised Learning. Clustering is a data set is divided into several groups or certain types of processes, and such that the object of the same group have a high degree of similarity, and the data between the different groups of similar objects is very small. Similar or dissimilar metrics are based on the value of the data object description to determine. Generally it is to use the distance between objects to be described. The group of physical or abstract objects, according to the degree of similarity between them, divided into several groups, which constitute a group of similar objects, this process is called clustering process, a cluster, also known as clusters, that is, by the set similar to each other a group of objects composed of different clustering object is usually not similar. Cluster analysis is a typical combinatorial optimization problems. Some commonly used for the respective individuals that have certain characteristics for classification. There are two main types of clustering algorithms: divide clustering and hierarchical clustering.

K-means Clustering [4] algorithm is a well-known partition clustering algorithm. Its main idea is given a set of data points and the required number of clusters K (K is specified by the user), K- means algorithm based on a distance function repeatedly the data points into K clusters in. At the beginning of the algorithm, the first K randomly selected data points as the initial cluster centers. Then calculate the distance between each data point and each seed cluster centers, assigning each data point to its distance from the nearest cluster center. Cluster center and allocated to it represents a cluster of data points. Once all the data points are assigned, the cluster centers of each cluster will be recalculated based on the existing cluster of data points. This process will be repeated until a termination condition is satisfied. The main advantage of K- means algorithm is simple and efficient, the disadvantage is that the data set can only be applied on the mean can be defined, and is very sensitive to outliers.

Hierarchical clustering is another major clustering methods. It generates a series of nested clustering tree to complete the cluster. Single point cluster (containing only one data point) in the bottom of the tree, at the top of the tree has a root node of the cluster. Root node covers all data points. Hierarchical clustering main merge (bottom-up) clustering and divisive (top-down) clustering in two ways. The main advantage of hierarchical clustering algorithm is its ability to use any form of distance and similarity function, the disadvantage is inefficient, because it requires computational complexity square.

Association Rules

Association rules [5] is to describe the potential relationship between the rule database data item, and associated data items that, according to the emergence of a transaction of some items, can be derived from other items in the same transaction also appeared, mining problems association rules can be divided into: discover the largest item sets and rules to generate two steps. The discovery is the largest project set of core mining association rules association rules mining algorithm, initially AISHE and SETM two algorithms, but they produce a lot of unnecessary candidate set during execution, the large amount of calculation. Thus the amount of data in association rule mining process is large, it is necessary to adopt some effective techniques to improve the efficiency of the algorithm. In addition you can use parallel technology to solve. In parallel algorithm involves coordination between computing, communications, memory utilization and the like. In fact, the valuable association rules often appear in the concept of a relatively high level, from a lower layer concept is difficult to find useful association rules. At present association rule has evolved from the concept of a single layer to the multi-layer concept, conceptual layer layers down from the general to the specific, association rules can provide more specific information, this is a gradual deepening of knowledge discovery process, become generalized association rules. Apriori association rule mining algorithm is the classical algorithm [6].

Apriori algorithm is one of the most influential mining Boolean association rules frequent itemsets algorithm (here, all the support is greater than the minimum support itemsets called frequent item sets, referred to as frequency set). Its core is based on a two-stage frequency set recursive algorithm thought. The basic idea of the algorithm is: first find all the frequent sets, these items appear frequently sets of at least a predefined minimum support same. Then set the frequency generated by the strong association rules, which must meet the minimum support and minimum confidence. Then use the set of rules found to produce the desired frequency, producing only contains all the rules of a collection of items, wherein each of the right side of a rule of only one, here is the definition of the rule. Once these rules are generated, then only those users more than the minimum rules given credibility was only to stay. To generate all frequencies set using recursive methods. However, it may produce a large number of candidate sets, and may need to repeat scan the database, are two major drawbacks of Apriori algorithm.

In addition to the classical algorithm, association rules Some research, for example: find language [7], weighted association rules algorithm [8] association rules, mining correlation and causation [9], the evolution of dynamic data mining association rules algorithm [10] to generate association rules different standard of measurement studies [11], a parallel discovery algorithm [10-12], and so on.

Conclusion

Data mining is a hot research era of big data, classification, clustering and association rules as a data mining method in the mainstream method of application of interest and the community, many of the proposed algorithm, each have their own advantages, only according to analysis of specific problems in the practical application, select the best algorithm for mining. This paper summarizes the common types of classification analysis of algorithms, some algorithms were analyzed, summed up their advantages and disadvantages. Various types of data mining algorithm combines the research achievements in various fields, with the development of the times, these theories will penetrate each other; mining methods will also be further developed.

Acknowledgment

The paper was supported by the project of excellent ICT engineer training program in Huanggang Normal University (Grant No. 2014zy04); and the project of zxfz2016A014 in Huanggang Normal University

References

[1] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques[M], 2nd ed. March 2006

[2] Quinlan J R. Learning efficient classification procedures and their application to chess and games ［C］/ /Michalski R S, Carbonell J G, Mitchell T M. Machine learning: an artificial intelligence approach, CA: Morgan Kaufmann, 1983:463-482.

[3] Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy(Centre for Distributed Systems and Software Engineering, Monash University).Mining Data Streams: A Review[C]. In SIGMOD 2005.

[4] Tian Zhang (IBM), Raghu Ramakrishnan (University of Wisconsin at Madison), Miron Livny (Universit of Wisconsin at Madison).BIRCH: An Efficient Data Clustering Method for Very Large Databases [C] Proceedings of the 2006 ACM SIGMOD.

[5] Rakesh Agrawal, Tomasz Imielinski, Arun Swami.Mining Association Rules between Sets of Items in Large Databases[C] Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993:63-82.

[6] Nan Jiang, Le Gruenwald (The University of Oklahoma, Norman). Research Issues in Data Stream Association Rule Mining[C], Proceedings of the 2006 ACM SIGMOD.

[7] Meo R, Psaila G, Ceri S.A new SQL-like operator for mining association rules [A]. Proc of the 22th International Conference on Very Large Database [C] .Bombay , India,1996.122- 133

[8] Cai C H, Fu W C, Cheng C H, et al.Mining association rules with weighted items[A] .IEEE International Conference on Database Engineering and Applications Symposium[C] , Cardiff, 1998

[9] Silverstein C, Brin S, Morwani R, et al.Scalable rechniques for mining causalstrucrures[A] .Proc 1998 International Conference on Very Large Data Bases [C] , New York, August 1998.594- 605

[10] Agrawal R.Parallel mining of association rules [J] .IEEE Transactions on Knowledge and Data Engineering, 1996, 8 (6):926- 969

[11] Park J S, Chen M S, Yu P S, et al.Efficient parallel data mining for association rules [A] .Proc Fourth Int' l Conf Information and Knowledge Management[C] .Baltimore, Nov 1995

[12] Cheung D W .Efficient mining of association rules in distributed databases [J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8 (6):910- 921

