Research on the Structured Data Mining Algorithm and the Applications on Machine Learning Field

Xiaodui Deng

Xi'an International University,
Xi'an, Shaanxi,710077, China

***Abstract*—In this paper, we conduct research on the structured data mining algorithm and applications on machine learning field. Various fields due to the advancement of informatization and digitization, a lot of multi-source and heterogeneous data distributed storage, in order to achieve the sharing, we must solve from the storage management to the interoperability of a series of mechanism, the method and implementation technology. Unstructured data does not have strict structure, therefore, compared with structured information that is more difficult to standardization, with management more difficult. According to these characteristics, the large capacity of unstructured data or using files separately store, is stored in the database index of similar pointer. Under this background, we propose the new idea on the structured data mining algorithm that is meaningful.**

Keywords- Data Mining, Structured, Applications, Machine Learning, Model Analysis.

Introduction

Database and information technology has been from the original file processing evolution to complex and powerful database system, which contains a large amount of data, and the data of the rich brought a claim to the powerful data analysis tools. Data mining is stored in a database, data warehouse, or the other information of a lot of interesting in the data mining process of knowledge [1-3].

From the perspective of traditional classification task, data mining techniques can be divided into seven categories, respectively is: classification, clustering, prediction, association rule, evaluation, visualization and the complex data type mining. But data mining technology obtained the swift and violent development in recent years has been extended to social network analysis, recommendation system, figure, spatio-temporal data analysis, data mining, feature selection, and so on with the new research fields. Data mining system framework is generally made up of three parts: data preparation system, modeling and mining system, the interpretation and evaluation system. (1) Data preparation is an important part in the process of data mining. If the data is ready to work well done, data quality is high, the process of data mining is more quick and convenient, finally dig out the patterns and rules to be more effective and applicable, the results are successful. (2) Data mining and modeling system is mainly based on mining theme and target, the algorithm and related technologies that analyze data, excavate the inner link between data and the potential rules. (3) The results of the data mining model concepts, rules, such as report chart form, if not satisfied with the results of mining, can be repeated in the previous step and link, until the concept of the ideal model of the rules or patterns.

From the intuitive understanding, machine learning is to make the machine to simulate the function of human learning. With the study of machine to simulate the behavior of the characteristics and ways of thinking to study, make it more intelligent machines as cannot only help mankind to do household chores, work, monitoring, etc. can further development to the machine like a human to learn, and we can also to human beings, by learning can have memory to absorb the knowledge they learned and can

through continuous learning to improve their performance. In the process of the information pattern recognition is the most important learning stage, and basic learning phase of feature extraction and classification rule acquisition as the top priority. In pattern recognition theory, the traditional used for feature selection methods have a lot of limitations, and mature cluster analysis method for its special properties used for feature selection has irreplaceable advantages. Use of a certain class separability measure a feature from the focus on selected the most conducive to the classification of feature subset known as the process of feature selection. After feature selection, the dimensions of the feature space also is compressed, it is more advantageous to choose most influential characteristics. The simplest method is to use the expert knowledge and experience and the most stringent method is under the condition of a given filtered by mathematical method as the feature extraction and feature selection in pattern recognition as two processes, their sequence is not fixed [4-5].

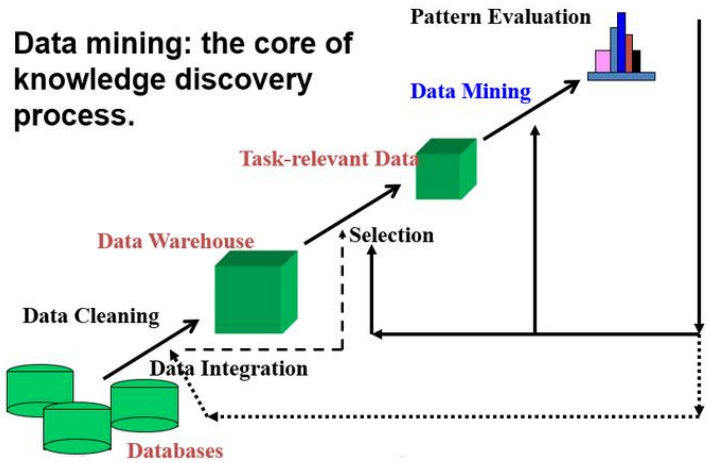

Figure 1. The General Primary Procedures of the Data Mining and Pattern Recognition

In this paper, we conduct research on the structured data mining algorithm and the applications on machine learning field. In the later section, we will discuss the issues in detail.

Our Proposed Methodology

The Machine Learning. Big data, machine learning, is not only a machine learning problem and the algorithm design, or a large-scale system problem. It is neither the simple machine learning is simple data processing technology can solve the problem, but a large and at the same time involved in the machine learning and data processing are two major aspects of its research subject.

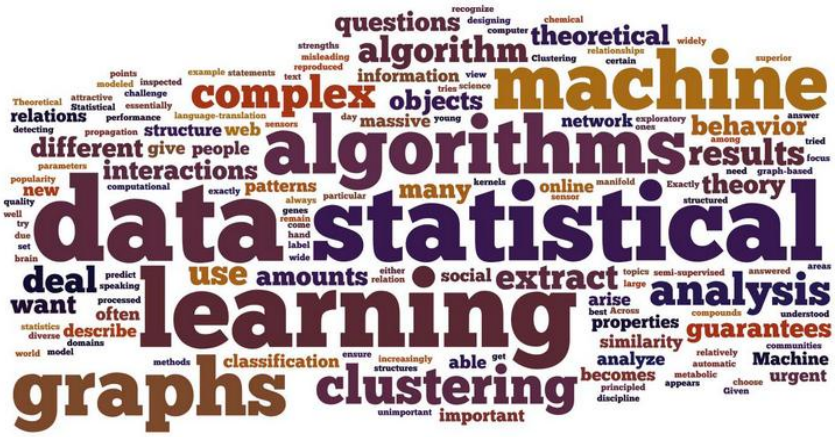

Figure 2. The Keywords of the Machine Learning

An ideal large data of machine learning system usually needs to have the following several aspects of the technical elements and characteristics. (1) It should provide a variety of the parallel training patterns, support different machine learning model and algorithm. (2) Need to provide abstraction of the underlying system, in order to realize to the underlying common data processing engine support, and provide scientific data in the commonly used programming language interfaces. (3) Should have open and rich ecology, wide application and ability of rapid evolution.

Now machine learning has become one of the main trends of intelligent, of course, people also use machine learning to solve many problems in life, it is also used in many aspects such as the object recognition in general human life field, optimal system, software engineering, network engineering. Relatively, we could summarize the application scenarios of the ML as the follows [6].

- Software defect prediction. Software engineering is a very important research topic that in the software defect prediction as it is based on software module on the basis of the defect record of past data, using the software in which module for now a flawed software defect prediction, in order to improve the performance of the software.
- Network traffic identification. Network use frequency more and more frequently, so we need to pay attention to the problem of the information security, which he introduced the network traffic identification method based on machine learning.
- Object recognition. Let the computer with the purpose of the visual ability to like people, it is need to identify the computer images and files, and classify the documents and the pictures according to certain rules and then the computer is the visual ability.

The Structured Data. With various types of data sources, information processing system need to different place and different time of data integration, data distribution process with the diversity of data requirements and become the more complex, data distribution not only needed to establish the relationships between data transmission that also need to provide data in time, space and function extension ability. The figure three shows the visualized structured data pattern [7].

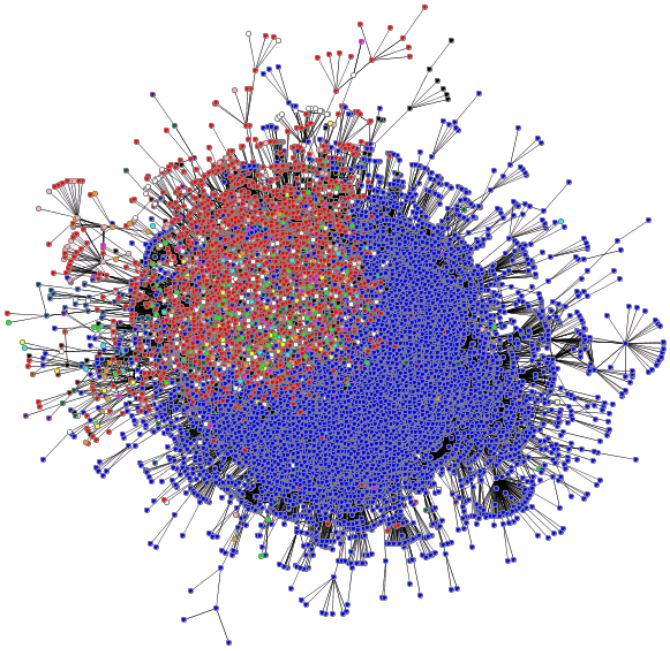

Figure 3. The Structured Data Pattern Distribution

Unstructured data must be corresponding interpretation software to open it and intuitive browsing, therefore, that cannot be directly obtained from the data itself is the expression of physical properties, which is not easy to get understand. Unstructured data and especially the multimedia data is very large amount of information, if stored in the database directly, in addition to greatly increase the capacity of the database and can reduce maintenance and application efficiency, particularly for the small and the medium-sized database system. The characteristics of the structured data could be summarized as the follows. (1) Processing speed is fast. The most striking feature is different from the traditional data mining, the generation of large data and update frequency is very fast, all have instant increase in per second, in the face of such a vast amounts of data, how to improve the efficiency of processing data, is an important topic of the enterprise. (2) Various data types. According to the various types, main can be divided into two types: the data structured data and unstructured data. Compared with the previous facilitate storage is given priority to with text structured data, unstructured data more and more. (3) Massive data as amount of data that cannot use existing technology to the management basically is to point to from dozens of TB to several PB, and with the progress of technology, of quantity increase.

The Pattern Recognition. To let the machine with pattern recognition ability, people need to first study of the human ability to recognize, so the mathematical model of pattern recognition is the study of human ability to recognize and with the basic aid of computer science and technology for computer simulation of human recognition behavior. Pattern recognition is to study how to make the machine, in other words, observe surrounding environment, identifying type of interest from the background, and the model classes belong to make accurate and rational judgments [8].

Pattern recognition is to point to the characterization of various forms of information processing and analysis, to describe things or phenomena, identify, classification and explanation of process as is an important part of information science and artificial intelligence. (1) Decision theory method. Also called statistical method, is the development of earlier and more

mature. Identified object first, digital transformation is suitable for computer processing of general digital information. A model is often represented with a large amount of information. Many pattern recognition systems in the digital link after preprocessing which is also used to remove the interference with information and reduce some deformation and distortion. (2) Syntactic methods that also called structure method or linguistics. Its basic idea is to put a model described as the combination of simpler sub-pattern, sub model and can be described as the combination of simpler sub-pattern, end up with a tree structure to describe, at the bottom of the simplest primitive sub-pattern called mode. In syntactic approach selected primitive problem is equivalent to select characteristic problems in general decision theory method. (3) Feature extraction is derived from the filter data useful information, from the many features to find out the most effective characteristics, in order to reduce difficulty of subsequent processing characteristics of filter after the necessary calculation, patterns are formed by feature selection and extraction of space.

The Further Development of the Data Mining. The data source of data mining with structured data, heterogeneous data and semi-structured data, at present mainly to mining the data in a relational database, for complex types of the multimedia data, such as the Web data mining has caused the researchers' interest. Data preprocessing is mainly consists of the general noise, incomplete, or not to technical specification of data processing and correction, including data cleaning, data smoothing and data merging, data conversion, etc. Data mining as a general process of human-computer interaction, repetition, the expert's knowledge or background knowledge in the field of the application with the complement and promote the process of mining, often used as a guide the discovery process to avoid meaningless result. In addition, the general method of data mining in the database content only on production rules, the rules is difficult to understand which can produce the knowledge or background knowledge in the field of application is easy to understand rules.

- New clustering method research. Clustering is a classic in the field of data mining tasks as data can be divided into different subsets, each subset with similar attributes or characteristics in the data. The K-means is a kind of the most classical clustering algorithm, but it still exist in the practice influenced by initialize clustering number, slow speed of clustering problem.
- Natural language processing research. Calculated by using the natural language processing is a method of dealing with the language semantics and other specialized technology. With the development of technology, more and more researchers began to study text classification, automatic abstract application problems, such as, which become an important application of the data mining research.
- Trust network research. In many e-commerce system recommendation systems, as well as the social network system, the use of trusted network technology can improve the credibility of the network society, which can be more reliable to the development of network application. In the trust network study, the core problem is the transmission of trust relationship, which in turn depends on the similarity measurement between different users.

Conclusion

In this paper, we conduct research on the structured data mining algorithm and the applications on the machine learning field. Traditional data mining algorithm as a general result of the data consistency constraints, in the management of the large-scale data sets, storage conditions, the local data in data update, failure,

and inefficient working system extensibility, etc. The solution is: through relaxing to the requirement of data consistency, cancel the associated complex queries, combined with specific application to improve the usability of the system. But due to the large number of records stored in the same table space, single table will reach billions of the billions of the records. Structured data refers to structure implied or no rules, no rigorous self-descriptive data. It is different from no structure of the file system, also do not have the structure of the database system is rigorous, but somewhere in the between. Under this background, we propose the new perspective and methodology on the mentioned issues that will be meaningful and innovative.

Acknowledgement

This research is financially supported by the planned project of technology bureau of Shaanxi province in 2014 (NO. 2014JM8356). The project name is: Research on the students learning state analysis and control based on intelligent Agent within the digital campus environment.

References

[1] Holzinger, Andreas, and Igor Jurisica. "Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions." Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Springer Berlin Heidelberg, 2014. 1-18.

[2] Mansmann, Svetlana, et al. "Discovering OLAP dimensions in semi-structured data." Information Systems 44 (2014): 120-133.

[3] Bellet, Aurélien, Amaury Habrard, and Marc Sebban. "A survey on metric learning for feature vectors and structured data." arXiv preprint arXiv:1306.6709 (2013).

[4] Cambria, Erik, et al. "Knowledge-based approaches to concept-level sentiment analysis." IEEE Intelligent Systems 2 (2013): 12-14.

[5] Paulheim, Heiko, and Christian Bizer. "Improving the quality of linked data using statistical distributions." International Journal on Semantic Web and Information Systems (IJSWIS) 10.2 (2014): 63-86.

[6] Nishanth, Kancherla Jonah, et al. "Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts." Expert Systems with Applications 39.12 (2012): 10583-10589.

[7] Moeyersoms, Julie, et al. "Comprehensible software fault and effort prediction: A data mining approach." Journal of Systems and Software in review (cit. on pp. iii, 134, 135) (2014).

[8] Jung, Kenneth, et al. "Functional evaluation of out-of-the-box text-mining tools for data-mining tasks." JAMIA. 22.1 (2014): 121-131.