Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/21/20231143

An overview of big data mining and data privacy protection technologies

Jinyang Liu

Wuhan City College, Wuhan, Hubei, China, 430075

jyliiu@163.com

Abstract. With the advent of the era of big data, data mining techniques have significantly improved their ability to extract valuable information from data. However, privacy dangers are growing. Consequently, securing the protection of personal privacy during the mining of massive amounts of data has become a significant challenge. This paper examines the relationship between data mining techniques and privacy protection measures through a review of the pertinent literature. It provides a concise analysis of the benefits and drawbacks of commonly utilized classification algorithms in data mining. In addition, it examines the interplay between data mining techniques and privacy protection and summarizes important privacy protection techniques. In addition, this paper provides a summary of the most important privacy protection methods. These techniques include data anonymization, association rule concealing, data perturbation, etc. By comprehending these privacy protection techniques, appropriate privacy safeguards can be selected to ensure the privacy and security of the data when conducting data mining.

Keywords: big data mining, privacy protection, big data processing workflow, decision tree algorithm, k-means algorithm.

1. Introduction

As the scope of human activities continues to expand, the volume of data has exploded. Diverse industries have amassed vast quantities of data, which are expanding in size and variety. This has surpassed the capabilities of conventional data management systems and processing methods, resulting in the emergence of the "Big Data" era. The "data" in Big Data is inherently authentic and trustworthy due to the underlying assumption of spatiotemporal homogeneity, which implies a consistent and interconnected relationship between all entities in time and space. This means that the attributes and patterns of any entity can be expressed via a unified digital signal so long as the proper encoding is applied. The extraction of useful information from immense amounts of data and the subsequent elucidation of business models and knowledge concealed within big data has become an important research field in the field of knowledge discovery. While utilizing data mining techniques to extract valuable information from large data sets, securing the security and privacy of the data has become a major challenge.

This paper utilized a literature review method to summarize and analyze prevalent algorithms and techniques in the disciplines of big data mining and privacy protection, providing readers with a comprehensive understanding. It helps readers comprehend the various data mining algorithms and

DOI: 10.54254/2755-2721/21/20231143

privacy protection measures, as well as their strengths and limitations. This paper also emphasizes the close relationship between data mining and privacy protection. In the process of data extraction, the protection of personal privacy is of the utmost importance. These summaries and analyses can provide valuable insights and references for researchers and professionals in related disciplines, enabling them to select appropriate methods and techniques for practical applications, thereby enhancing the efficacy and effectiveness of data mining and privacy protection.

2. Workflow for big data processing

The following components make up the bulk of the big data processing workflow: First, the big data collecting. Data is collected from a wide variety of smart hardware devices, sensors, online pages, mobile app applications, etc., and stored in many databases. The data in these databases can also be processed in the most fundamental ways. Web scraping, log collection, data collection from mobile apps, and automatic information collection from IoT (Internet of Things) devices are all examples of popular approaches to gathering data; Second, we must import and preprocess the data. Multiple databases may be present in the end of data collection, but importing the collected data into a centralized large-scale distributed database or distributed storage cluster is recommended for efficient analysis. Data cleaning and preparation are done concurrently on the basis of data import. Data cleaning is performed with the intention of accomplishing certain tasks, such as standardizing the data format, removing outliers, correcting errors, and getting rid of duplicates. Integrating data is bringing together disparate data sets and storing them in one central location, or "data warehouse." The data is transformed using procedures like smoothing aggregation, data generalization, and standardization so that it can be used in data mining. Finally, data reduction is performed by data summarization to identify useful features that depend on the discovery purpose, hence lowering the data size and requiring a minimum amount of data. Third, administration and archiving. Technology advancements like distributed file systems, NoSQL databases, and cloud storage have emerged as primary solutions to the storage and distributed storage needs of big data's massive volumes of semi-structured and unstructured data. Analysis and statistics are the fourth point. At this point, it is usual practice to utilize distributed databases or distributed computing clusters to analyze and aggregate the huge amounts of data contained in them. Last one is big data mining. Data mining is the process of developing these models through a combination of exploratory and computational techniques. The process begins with collecting and organizing data and continues with the development of models based on the findings.

3. Commonly used algorithms in data mining

There are now four approaches utilized in the data mining phase: Naive Bayes, Support Vector Machines (SVM), AdaBoost, Decision Trees, and other classification methods are just the beginning. Algorithms like BIRCH, K-Means, Expectation-Maximization, and K-Nearest Neighbors (KNN) are used in clustering, which is the second topic. As a third category, we have association rules, which encompasses a wide variety of algorithms. Last but not least, there is predictive modeling, which makes use of methods like Sequential Pattern Mining for Generalized Clusters (SPMGC), Regression Models, and others.

3.1. Decision tree classification algorithm

One popular machine learning method dealing with classification issues is the Decision Tree Classification Algorithm. It builds a tree-like structure to organize data sets. The nodes of a decision tree are referred to as the "root," "internal," and "leaf" respectively. The top level, or "leaf," node reflects the final classification findings, whereas the bottom level, or "root," node represents the initial feature.
The primary concept behind the decision tree classification algorithm is to recursively partition the dataset so that the data in each partition is highly correlated with one another. When building a decision tree, the algorithm chooses the most informative split criteria and assesses the quality of the split based on how clean the resulting subgroups are. Several popular decision tree algorithms exist today [1], such as ID3, C4.5, and CART. The benefits of using the decision tree classification method are as follows:

DOI: 10.54254/2755-2721/21/20231143

The first advantage is that it is straightforward and easy to grasp. For another, it's capable of dealing with issues involving multiple classes at once. Third, it accommodates both quantitative and qualitative characteristics. The fourth flaw is that it ignores extreme values. Nonetheless, there are constraints associated with the decision tree classification algorithm: As a first issue, it tends to overfit, which is especially problematic when working with complicated and high-dimensional datasets. Secondly, it is highly sensitive to initial conditions, as even little changes in the input data can result in drastically different tree architectures. Thirdly, it has trouble with continuous variables and missing value situations.

3.2. K-means algorithm

The K-means algorithm is a frequently employed unsupervised learning algorithm that divides a dataset into K distinct clusters or groups. Its purpose is to divide the sample points into clusters with a high degree of similarity within each cluster and a low degree of similarity between clusters.

The K-means algorithm's benefits include its simplicity and ease of implementation, as well as its high computational efficiency and robust scalability. Nevertheless, the K-means algorithm has some limitations. For instance, the algorithm's results are extremely reliant on the initial selection of cluster centroids. Poor initial centroids selection can result in suboptimal outcomes. Therefore, substantial effort is necessary to effectively determine the initial centroid values [2]. Several variants of the K-means algorithm, including K-means++, K-medoids, hierarchical K-means, and others, have been developed to resolve some of its limitations. These algorithms seek to enhance the initial selection of cluster centroids or the determination of the number of clusters in order to achieve superior clustering outcomes.

3.3. Association rules algorithm

Association Rules is a data mining algorithm used to identify frequent item sets and associations within a dataset. It is based on transactional data within the dataset and derives meaningful rules by analyzing co-occurrence relationships between item collections. The most important algorithm for association rule mining is the Apriori algorithm, a highly influential algorithm for discovering frequent item sets and generating single-dimensional, single-level, Boolean association rules [3]. The association rule algorithm has numerous practical applications, including market basket analysis, recommendation systems, and cross-selling analysis. By identifying associations within a dataset, it assists individuals in comprehending the relationships between products or transactions, thereby providing decision support and business insights. Traditional methods for assessing financial risk in e-commerce frequently suffer from low accuracy, complex algorithms, and lengthy evaluation cycles, and fail to meet the urgent risk assessment requirements of businesses. The association rule algorithm-based e-commerce financial risk assessment system manages and maintains e-commerce financial data via the data collection module and transmission control module. It builds a risk indicator system using financial indicators such as investment risk, operational risk, asset value, liquidity risk, and internal/external factors in order to evaluate the risk of e-commerce financial data. As depicted in Figure 1, when predicting the risk values of the same set of financial data, various risk assessment systems exhibit significantly higher prediction error probabilities than the traditional risk assessment system, and the variability is unstable. In contrast, the e-commerce financial risk assessment system based on the association rule algorithm exhibits significantly lower prediction error probabilities than the conventional assessment system, with the maximum error probability not exceeding 10%.

DOI: 10.54254/2755-2721/21/20231143

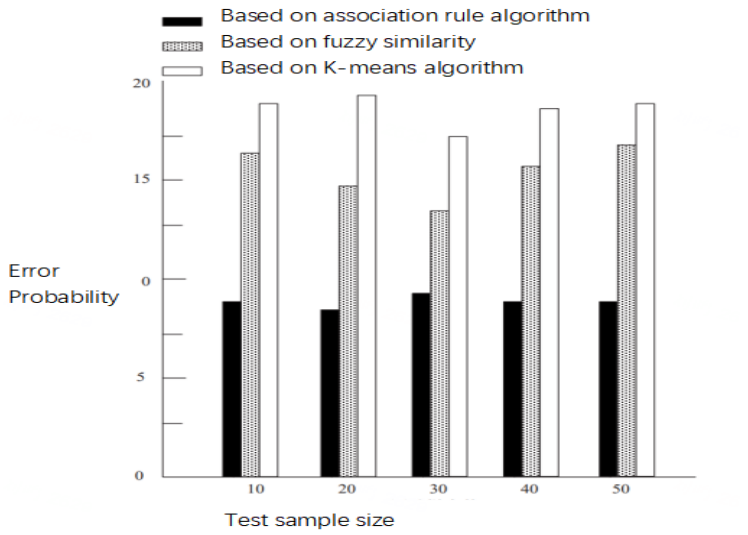

Figure 1. The comparison of error probabilities·2022 [4].

4. Data mining privacy protection techniques

As the applications of data analytics continue to expand, the need to safeguard the privacy of individuals during the monitoring process has become more pressing. There is a growing demand for privacy data extraction and analysis that is both meaningful and protects privacy. Privacy protection has become a crucial concern that must be addressed in the process of data analytics [5].

4.1. Data anonymization

Anonymization technology is a prevalent privacy protection method. It involves modifying, generalizing, or adding noise to data in order to conceal sensitive information and preserve user privacy. Common methods of data anonymization include, but are not limited to: First, k-anonymization, which entails generalizing and perturbing the data such that each record in the dataset is indistinguishable from at least k-1 other records based on their attributes, thereby concealing individual identities. Second, the l-diversity model is predominantly concerned with enforcing the diversity of sensitive attributes within each equivalence class. It guarantees that sensitive attributes in any equivalence class have at least l distinct values, where l is the number of distinct values for the sensitive attribute within the equivalence class [6].

4.2. Association rule hiding

Association rule concealing is a data mining technique employed for the protection of privacy. Its objective is to identify beneficial association rules within a dataset while protecting individual privacy. There are two primary categories of association rule concealing techniques: distortion and blocking [7]. The purpose of this method is to ensure the complete protection of individual privacy while preserving data accessibility and analytical efficacy. The transformation method for privacy protection of association rules involves transforming or perturbing the original association rules to safeguard sensitive information and individual privacy. First is item transformation, followed by value transformation, structure transformation, support and confidence transformation, and rule selection. The concealing methods do not modify the sensitive rule data; rather, they conceal the frequent item sets that would generate sensitive rules [8].

5. Privacy protection technology based on data distortion

It is a common method for preserving data privacy that involves intentionally distorting or transforming the original data in order to conceal sensitive information and protect individual privacy. Although these

DOI: 10.54254/2755-2721/21/20231143

techniques can reduce the precision of data, they still provide a certain level of data accessibility and utility. Currently, techniques for data distortion-based privacy protection include randomization, barring, and aggregation [9].

Random perturbation is a technique that adds randomness to a dataset. It can be accomplished by arbitrarily permuting the order of data records or by randomly replacing or exchanging attribute values in order to disrupt the data structure. This prevents the inference of particular identities and relationships. Random sampling is the process of selecting samples from a dataset at random. It can be used to generate a sampled dataset with comparable statistical characteristics, thereby reducing reliance on the original data while protecting individual privacy. In blocking, the hidden information is substituted with ambiguous symbols or placeholders. It is possible to determine a range of minimum and maximum approximated values by counting the number of uncertain symbols. This enables the processing and analysis of data while maintaining confidentiality. The aggregation method requires the grouping of data attributes to ensure that no two distinct records exist within the same group.

6. Conclusion

This paper provides a concise analysis of three aspects: the process of big data processing, commonly employed data mining algorithms, and data mining privacy protection techniques. In addition, some common data mining algorithms and solutions for protecting privacy are presented. In addition, it examines the advantages and disadvantages of specific algorithms and solutions. In the context of data mining, this paper examines commonly employed algorithms for classification, clustering, and association rule mining. Regarding privacy protection, the article investigates the importance of privacy protection and its application in data mining. In addition, it introduces a number of privacy protection techniques, including K-anonymity, association rule concealing, and data distortion. Due to the exponential development of data volume and complexity, it has become imperative to develop more efficient and precise data mining algorithms to manage massive datasets. In addition, there is substantial space for improvement by incorporating cutting-edge technologies such as machine learning and deep learning into data mining algorithms. This integration can further improve the precision and effectiveness of data mining operations. Moreover, as the risk of data privacy intrusions rises, ongoing research and innovation in privacy protection techniques are required. The investigation of advanced privacy protection models and algorithms, such as differential privacy and homomorphic encryption, can result in more robust privacy protection capabilities. On the other hand, data mining and privacy protection have implications for numerous industries and domains, including healthcare, finance, and social networks. Future research could concentrate on in-depth exploration tailored to the particular requirements of these domains, resulting in customized data mining and privacy protection solutions. This targeted approach would address the specific challenges and requirements of each industry, resulting in more effective and efficient data analysis while preserving confidentiality.

References

[1] YIN Tingjun, LI Linghui, Zhou Rui.Overview of Data Classification Algorithms for Big Data Mining[J].Digital Technology &Application, 2021,39(01):103.DOI:10.19695/j.cnki.cn12-1369.2021.01.32.

[2] YAO Qifeng, YANG Lianhe, Research on Classical Classification and Clustering Algorithms in Data Mining[J].Modern Information Technology,2019,3(24):88.DOI:10.19850/j.cnki.2096-4706.2019.24.029.

[3] Yang Xiaojuan. A review of domestic research on data mining[J]. Computer Programming Techniques and Maintenance,2020,No.422(08):116.DOI:10.16184/j.cnki.comprg.2020.08.041.

[4] Figure1·The comparison of error probabilities 2022·https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7ioT0BO4yQ4m_mOgeS2ml3UIm3JTwcW8QPppm56w26gKeqat7MPhwKZoL2BcEFMO_&uniplatform=NZKPT

DOI: 10.54254/2755-2721/21/20231143

[5] SU Pengchong, YUAN Deyu, MA Ding,Research on the Big Data Mining Technology Based on Privacy Protection[J].Modern Computer (Professional Edition),2017(20):26.
[6] Zhang Bo,Privacy-Preserving Data Mining Analysis[J].INFORMATION & COMMUNICATIONS,2018,No.191(11):172.
[7] Aggarwal Charu C,S Yu Philip. Privacy-preserving data mining :models and algorithms [M]. Springer Science & Business Media,2008
[8] Yang Yang, Chen Hongjun, A review of Research on Privacy Preserving Data Mining Technology[J].Microcomputer Applications,,2020,36(08):43.
[9] Si Ruoqian.Research and Application on Privacy Protection for Data Mining[D].Nanjing University of Posts and Telecommunications,2016:9.