ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940www.ijsem.org

Vol 11, Issuse.4 Dec 2019

A Thorough Study and Examination of Large Data Employing Data Mining Methods

P Chandra Sekhar Reddy,Dr. Suraj V Pote

Abstract: Contemporary data management systems search through enormous datasets to find patterns and correlations that were previously undiscovered in addition to storing and retrieving data. The need for computer applications and data mining software is rising as a result of how quickly new technologies are being developed. To ensure that all calculations lead to the same result, the necessary software and tools need to work with remote databases. However, because of legal restrictions and the necessity for a competitive edge, distributed data mining presents privacy concerns. Experts in the fields of big data, cyber security, and data mining are therefore motivated to study more. Researchers created Privacy-preserving Distributed Data Mining (PPDDM) to address the multi-party computation problem, in which multiple users attempt to perform a data mining task cooperatively using their respective private data sets, in order to get around these limitations and benefit from these advantages. Participants discover only the outcomes of the data mining method and their own inputs after finishing the exercise. The main goal of this study was to develop a novel way to privacy-preserving data mining for the purpose of developing Decision Tree Classifiers using vertically partitioned data. Weka is utilized to construct a conclusion tree classifier using the proposed PPDM algorithm, and the outcomes are contrasted with the well-researched J48 approach. This analysis employs accuracy and precision as its standards. Compared to the conventional approach, the suggested PPDM algorithm offers far greater accuracy and precision.

Keywords: Big Data, Data Mining, PPDM,PPDDM, Etc.

I. Introduction

The process of searching through enormous amounts of historical datasets for relationships and insightful information that have not yet been found is known as data mining. This method requires the exploration of big datasets in order to identify and analyze patterns. The successful extraction of patterns from data necessitates applying sophisticated techniques. This particular topic is under the purview of computer science, which draws upon a broad range of notions and theories from numerous other academic disciplines to meet the difficulties it encounters. Though its most basic definition can also be described as "data mining," it is commonly referred to as

1Research Scholar, Department of Computer Science Engineering, University of Technology, Jaipur
2 Professor, Department of Computer Science Engineering, University of Technology, Jaipur

"data mining." The process of extracting valuable insights from massive amounts of unstructured data is known as data mining. What is known as "big data analytics" is the methodical investigation of data patterns found inside enormous databases using one or more computer systems. Often used in the context of data mining is the abbreviation KDD. KDD is an acronym for "Knowledge Discovery in Data." A method known as "data mining" is used to sift through massive datasets kept in databases in search of meaningful patterns or relationships. You might find yourself needing to invest a significant amount of extra time in this method. "Data mining" is defined by the authors of a publication titled "Data Mining: A Comprehensive Overview" (Lausch et al., 2014) as the time-consuming process of finding significant and potentially helpful patterns inside databases. To be more precise, they call data mining "the complex process of finding meaningful and possibly advantageous patterns within databases." It takes a lot of work to extract meaningful insights from ambiguous data, which is a challenging undertaking. "Data mining" is the process of extracting relevant information or knowledge from sizable datasets or databases. The discipline of computer science encompasses this topic of study. Pattern recognition, machine learning, and statistical analysis are a few examples of the computational techniques that are employed. According to Lindell et al. (2000), data mining is the process of methodically extracting valuable data from large datasets through the use of computational techniques. The following categories apply to the two components.

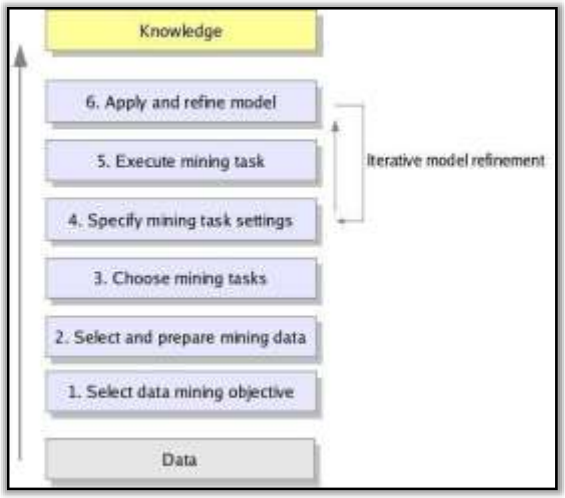
Knowledge
6. Apply and refine model
5. Execute mining task
4. Specify mining task settings
3. Choose mining tasks
2. Select and prepare mining data
1. Select data mining objective
Data
Iterative model refinement

Figure 1: Procedure for Mining Data

In order to store the models and other artifacts created throughout the data mining process, a data mining system must contain both a data mining engine and a repository. The system won't be able to operate as effectively as it can till then. Vikas et al. (2011) claim that important information and insights may be extracted from big databases via a process called data mining. For your viewing enjoyment, a graphical depiction of this concept is provided in Figure 1.1.

II. Literature Survey

Zhong et al. (2007) presented the Guided Perturbation method as a fresh approach to PPDM Perturbation. The previously mentioned technique demonstrates substantially faster performance than the cryptographic method while maintaining a similar level of accuracy.

The authors of the work by Weiwei Fang et al. (2008) have developed a novel approach to decision tree training that guarantees the privacy of users. The outcomes of the experiment demonstrate how well the method works to maintain anonymity while maintaining accuracy and effectiveness.

Homomorphic encryption-based decision tree approach was first presented by Fang et al. in 2009. The data's efficacy, integrity, secrecy, and honesty were all successfully protected by the used methodology.

Shen et al. (2009) used a somewhat trustworthy third party to develop the PPDM decision tree method, PPC4.5, to improve collaborative computation.

For horizontally partitioned data, Jhalla et al. (2016) proposed a privacy-preserving data mining (PPDM) approach that utilizes perturbation techniques and linear transformations such as the Walsh-Hadamard Transform (WHT). The studies made use of the Iris and WDBC datasets in their final forms. The results of many linear transformations were examined using Weka. The results showed that the suggested method achieved precisions that were on par with the K-NN classifier.

III. Experimental Evaluation of the Proposed PPDM Decision Tree Algorithm

the process of classifying every single data piece into a predefined category. Every item is categorized into a number of predefined groups. In the fields of data analysis and machine learning, a classifier is built to forecast attributes linked to a particular class label. In the field of data mining, categorization is essential since it makes it easier to arrange data into meaningful and cogent groups. Making accurate predictions about the classification of individual items within a given collection is the primary goal of categorization. Borrowers can be categorized into low, medium, and high credit risk groups using a categorization model. Classification is a commonly used technique in several fields, such as credit analysis, medicine response modeling, market research, product development, and customer division.Among the data mining classification techniques are: • Decision tree induction techniques

- Assist vector devices

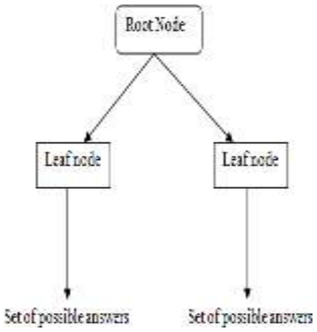

Fig 2: Decision Tree

Among these are rule-based approaches, neural networks, Bayesian networks, and memory-based learning. The Decision Tree technique is widely used in data mining and is well-known. The fundamental objective of a decision tree is to create a predictive model that can estimate the value of a goal modifiable from a set of input factors. In Figure 4.1, the core nodes represent the input variables, and the edges to their child nodes represent potential values for those variables.

The algorithms that are used to induce decision trees are recursive. To get started, let's define a first characteristic. Partitioning the data is essential for the tree to retain its compact structure. in a productive way at the root node. The goal is to reduce the quantity of cases) that require attention. It won't be possible to classify every thing until they are all uniformly classed. A divide may be deemed to be the most ideal if it offers the possibility of improving comprehension.

IV. Methodology Used

The data is stored at many places or locations. During testing, the data is divided in half vertically, with the equal amount of samples in each half. A distinct collection of traits sets each of these categories apart from the others. The privacy of the information kept in the shared database is something that both parties have an interest in maintaining. Throughout the decision tree building process, the data identities of both parties must be kept confidential. WEKA 3.8 implements the proposed Privacy-Preserving Data Mining (PPDM) technique. To construct a decision tree, the following actions must be taken: To decide how to divide up the qualities, careful math is done. Utilizing the most effective technique for forming subgroups.

The experimental data was divided into categories using entropy and the Gini index. Between the two sets, there are 351 instances that are similar, and a total of 17 attributes (including the class attribute). Since the entropy of each entity is influenced by the presence or absence of class attributes, it is computed separately for each entity. Each side's perception of the value of its advantages will vary depending on the particular circumstances surrounding it. Every participant had equal access to the gain values, and the hierarchy's base was determined by the most valued attribute. All parties to this transaction are aware of the profit as the only value. This approach satisfies the algorithm's need for anonymity. It is feasible to build the tree. in a comparable manner keeping all personal information secret. The tree therefore becomes a very useful tool for both of them. The algorithm utilizes secure multiparty computation (SMC) and scalar product protocols.

V. Results

The trials are carried out in an environment that simulates a distributed system. To simulate a vertical split, an arbitrary division of the dataset has been made. To execute the tests, the dataset is divided into two equal halves along the vertical axis. First, the J48 method in Weka—a software implementation of the C4.5 algorithm—is used to build a decision tree. The development of this decision tree is predicated on the idea that all of the data is contained in one single area and isn't divided in any way. The results of the current technique are compared with the algorithm that is offered for the vertical partitioning of data. The program takes into account knowledge acquired and the best split technique for choosing how to divide the data appropriately. In addition, the process is designed to maintain the consistency of the counterpart's data while creating a tree structure out of the combined data.

Table 1: Accuracy

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy
J48	321	30	91.453%
Proposed PPDM Algorithm	346	5	98.5755%

Table 2: Precision

Algorithm	TP Rate	FP Rate	Precision
J48	0.915	0.125	0.915

Proposed PPDM Algorithm	0.986	0.022	0.986

The true positive rate of 0.986 for the suggested algorithm and the false positive rate of 0.915 for J48 indicate that the former has a lower predictive accuracy than the latter. Additionally, the recommended PPDM algorithm has a lower false positive rate than J48, which indicates that it produces fewer incorrect predictions. The outcomes of both strategies in terms of the TP Rate and the FP Rate are displayed in Figure 4.9. The graphic demonstrates that the suggested technique obtains a higher True Positive (TP) Rate in comparison to the baseline J48 algorithm.

VI. Summary

The proposed method was applied to a real-world dataset to construct the decision tree. A centralized database is used by the well-known data mining program Weka 3.8, whose output tree is compared to ours in order to assess the validity and correctness of the tree. The gathered data demonstrates that the intended approach works significantly better than the default algorithm, J48. The proposed method performs significantly better than the current J48 algorithm since it can be applied in a distributed scenario and safeguard the privacy of the different parties.

REFERENCES

[1]. Koufakou, A., Secretan, J., Reeder, J., Cardona, K., & Georgiopoulos, M. (2008, June). Fast parallel outlier detection for categorical datasets using MapReduce. In Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on (pp. 3298-3304). IEEE.

[2]. Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2009, April). Outlier detection in axis-parallel subspaces of high dimensional data. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 831-838). Springer, Berlin, Heidelberg.

[3]. Kriegel, H. P., Kroger, P., Schubert, E., & Zimek, A. (2011, April). Interpreting and unifying outlier scores. In Proceedings of the 2011 SIAM International

[4]. Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. Advances in engineering software, 69, 46-61.

[5]. Mirjalili, S. (2015). The ant lion optimizer. Advances in Engineering Software, 83, 80-98.

[6]. Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. Advances in Engineering Software, 95, 51-67.

[7]. Mohemmed, A. W., Zhang, M., & Browne, W. N. (2010, July). Particle swarm optimisation for outlier detection. In Proceedings of the 12th annual conference on Genetic and evolutionary computation (pp. 83-84). ACM.

[8]. Moonesinghe, H. D. K., & Tan, P. N. (2008). Outrank: a graph-based outlier detection framework using random walk. International Journal on Artificial Intelligence Tools, 17(01), 19-36.

[9]. Morales, G. D. F., & Bifet, A. (2015). SAMOA: scalable advanced massive online analysis. Journal of Machine Learning Research, 16(1), 149-153.

[10]. Nguyen, H. V., Ang, H. H., & Gopalkrishnan, V. (2010, April). Mining outliers with ensemble of heterogeneous detectors on random subspaces. In International Conference on Database Systems for Advanced Applications (pp. 368-383). Springer, Berlin, Heidelberg.