A COMPREHENSIVE ANALYSIS AND INSPECTION OF MASSIVE DATA USING DATA MINING TECHNIQUES

P Chandra Sekhar Reddy, Dr. Suraj V Pote
[1]Research Scholar, Department of Computer Science Engineering, University of Technology, Jaipur
[2]Professor, Department of Computer Science Engineering, University of Technology, Jaipur

Abstract: In addition to storing and retrieving data, modern data management systems sift through massive datasets to uncover patterns and correlations that were previously unknown. Because new technologies are developed so quickly, there is an increasing demand for computer applications and data mining tools. The required tools and software must be able to interact with remote databases in order to guarantee that every calculation yields the same result. Distributed data mining raises privacy concerns, nevertheless, due to regulatory limitations and the need for a competitive advantage. This encourages experts in the domains of big data, cyber security, and data mining to do more research.

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. Weak 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. Safe for privacy with the use of modern big data mining tools, data mining is possible. To determine which Big Data Mining Tool is the best, we examine several possibilities. The benefits and drawbacks of each instrument are compared to one another. Using Decision Tree Classifier, we experimentally assess these three methods on two datasets from the online UCI Repository.

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

I. INTRODUCTION

Data mining is the process of looking for links and useful information that have not yet been discovered by sifting through massive volumes of historical datasets. Large datasets must be explored in order to find and examine trends using this strategy. In order to successfully extract patterns from data, complex procedures must be used. This specific subject is within the domain of computer science, which, in order to overcome its challenges, incorporates ideas and theories from many different academic fields. It is usually referred to as "data mining," even if its most basic description is similarly comparable to "data mining." Data mining is the process of gleaning insightful information from vast volumes of unstructured data. "Big data analytics" refers to the methodical use of one or more computer systems to analyze data patterns discovered within massive datasets. KDD is a frequently used acronym in the domain of data mining. AKA stands for "Knowledge Discovery in Data." Massive datasets stored in databases are combed through using a technique called "data mining" to look for significant patterns or relationships. You may find that using this strategy requires a substantial amount of additional time. The authors of "Data Mining: A Comprehensive Overview" (Lausch et al., 2014) define "data mining" as the laborious process of identifying meaningful and perhaps useful patterns inside databases. Data mining, to put it more precisely, is defined as "the complex process of finding meaningful and possibly advantageous patterns within databases." Extracting actionable insights from vague data is a difficult task that requires a lot of work. "Data mining" is the process of drawing pertinent knowledge or information out of large databases or datasets. The field of computer science includes this area of research. Among the computer approaches used are statistical analysis, machine learning, and pattern recognition. Data mining, according to Lindell et al. (2000), is the methodical process of using computational tools to extract meaningful data from big datasets. The two components fall under the following categories.

II. LITERATURE SURVEY

Using safe multi-party calculations, the authors of the research [Animesh Tripathy et al., 2012] presented a classification technique for Privacy-Preserving Data Mining (PPDM). The procedure of chopping trees is found to have the capacity to improve both accuracy and covertness.

Two approaches were presented in a paper by Jinfei Liu et al. (2012) to improve the DBSCAN clustering's privacy guarantees. The techniques used have successfully raised the degree of secrecy about the different data categories.

In Rosa Karimi Adl et al. (2012)'s work, the researchers used an anonymization technique to build a simulated game with the goal of establishing mutually acceptable levels of privacy protection. The novel approach disclosed a series of games as a covert intelligence gathering tactic. A backward induction methodology was employed to analyze the complete symmetry of the game.

In order to create a result tree, Chahal (2013) suggested an enhanced version of the ID3 technique. The author also used a real-world dataset to demonstrate how to use this improved approach. The application of substitute values obtained from the tree enabled the development of well-informed approximations concerning the value of the class attribute.

Hussain et al. unveiled a revolutionary data mining methodology in 2014 with the goal of protecting privacy when analyzing big datasets. To achieve the intended result, this study used cluster analysis and cryptography techniques. A rule-based methodology was used to carry out the clustering procedure. For every dataset, a separate set of criteria was used. A new evaluation process has been unveiled, consisting of three separate levels: private, authorized, and open.

In a research paper by Nasrin Irshad Hussain et al. (2014), a new approach to protecting user privacy inside large datasets using cryptographic techniques including encryption and key management was presented. A rule-based methodology was used to carry out the clustering procedure.

Jhalla et al. (2016) offered a privacy-preserving data mining (PPDM) method for horizontally partitioned data that takes use of linear transformations like the Walsh-Hadamard Transform (WHT) and perturbation methods. 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. METHODOLOGY USED

The information is kept in multiple locations. The data is split in half vertically during testing, with an equal quantity of samples in both halves. Every one of these categories is unique from the others due to a certain set of characteristics. It is in everyone's best interests to preserve the privacy of the data stored in the shared database. The confidentiality of the data identities of both parties is crucial during the decision tree building process. The suggested Privacy-Preserving Data Mining (PPDM) method is implemented in WEKA 3.**8.** In order to build a decision tree, the subsequent steps need to be completed: Careful calculation is needed to determine how to split up the attributes. Completed. Making use of the best method for creating subgroups.

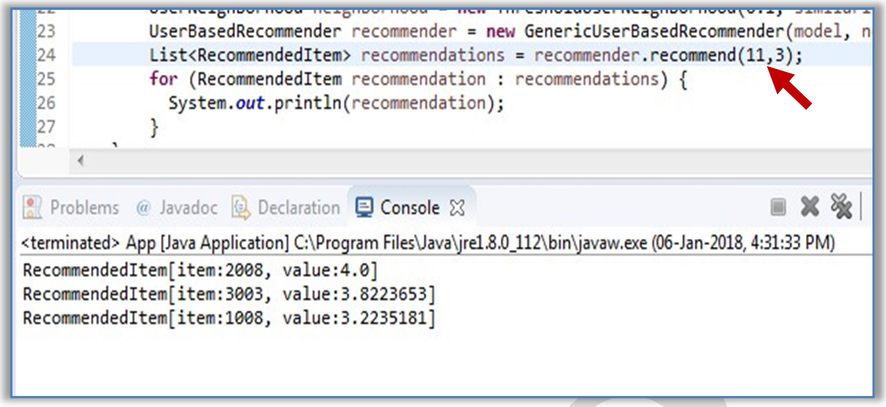
The Gini index and entropy were used to categorize the experimental data. There are 351 between the two sets. 17 attributes in total, including the class attribute, and instances that are comparable to each other. Since the existence or lack of class attributes affects each entity's entropy, it is calculated individually for each entity. The relative merits of each side's advantages will be viewed differently based on the specific conditions surrounding the situation. The highest valued attribute decided the base of the hierarchy, and each member had equal access to the gain values. Everybody involved in this deal understands that the profit is the only asset. The anonymity requirement of the algorithm is met by this method. Building the tree is a doable project. in a similar way, maintaining the privacy of all personal data. As a result, the tree turns into a really helpful tool for them both. Scalar product protocols and secure multiparty computation (SMC) are used by the approach.

IV. PROPOSED WORK

The purpose of this study is to forecast future demand for fastener goods by analyzing user input. In the manufacturing sector, a variety of factors are considered while making forecasts. The predictive power of the study is based on a suggestion system. Both the selling unit and the type of fastener that the consumer wants are tracked. A user-item matrix is made to determine the relationship between an item and a person. A Pearson correlation similarity metric is utilized to ascertain the level of similarity between clients. To predict the preferences of customers who have never used a specific product before, a novel data model has been developed. By depending on the suggestions or forecasts produced by the model, manufacturers can estimate the demand for their goods.

V. EXPERIMENTAL EVALUATION AND RESULTS

Users receive customized product recommendations from the recommender system based on their own interests and preferences. During the trial period, Customer 11 receives three recommendations from the recommender system. Figures 6.7 and 6.8 show how 11 and 15, respectively, were produced.

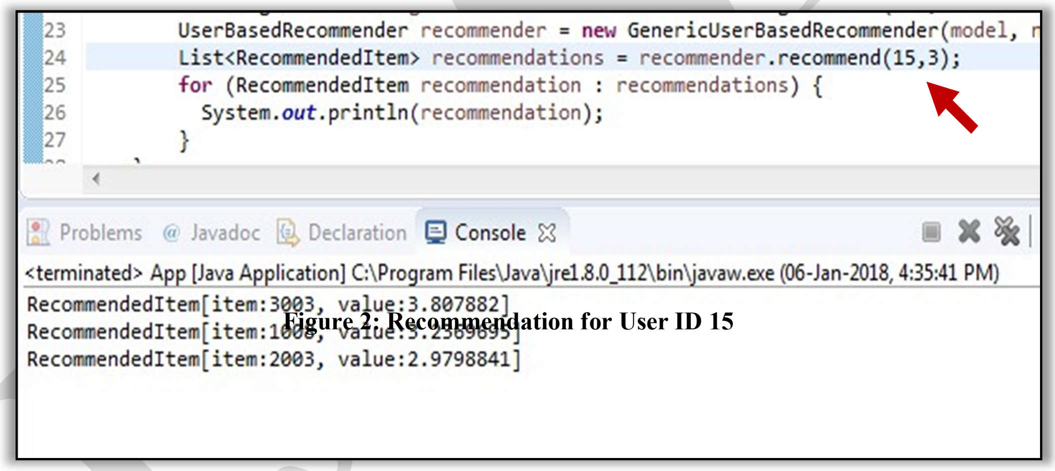


```
        UserBasedRecommender recommender = new GenericUserBasedRecommender(model, n
        List<RecommendedItem> recommendations = recommender.recommend(11,3);
        for (RecommendedItem recommendation : recommendations) {
          System.out.println(recommendation);
        }
```

```
<terminated> App [Java Application] C:\Program Files\Java\jre1.8.0_112\bin\javaw.exe (06-Jan-2018, 4:31:33 PM)
RecommendedItem[item:2008, value:4.0]
RecommendedItem[item:3003, value:3.8223653]
RecommendedItem[item:1008, value:3.2235181]
```

Figure 1: Recommendation for User ID 11

Product number 2008, which has a recommendation rating of 4, is advised that User 11 buy. It is advised that User 11, Bajaj Auto Ltd., take into consideration purchasing nuts from Sundaram Fasteners Ltd. The product's high recommendation value of 4 further supports the model's endorsement of the particular maker. The next two entries on the list are 3003 and 1008, with recommendation values of 3.9 and 3.2, respectively.



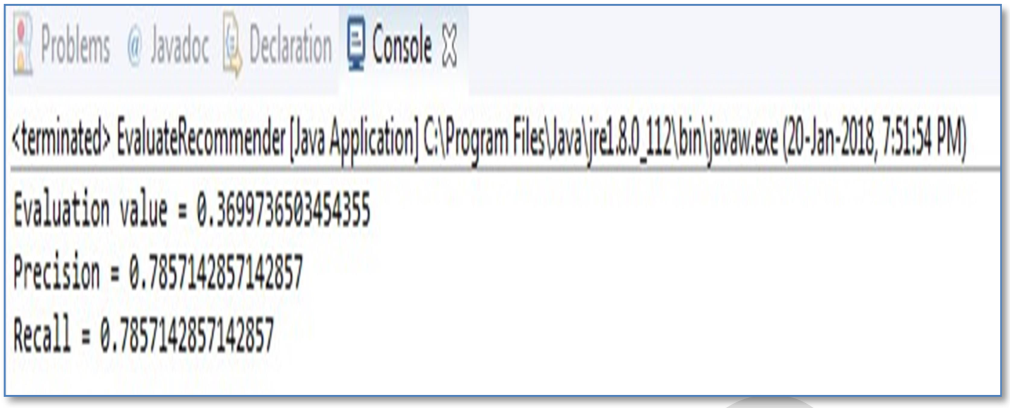
```
        UserBasedRecommender recommender = new GenericUserBasedRecommender(model, r
        List<RecommendedItem> recommendations = recommender.recommend(15,3);
        for (RecommendedItem recommendation : recommendations) {
          System.out.println(recommendation);
        }
```

```
<terminated> App [Java Application] C:\Program Files\Java\jre1.8.0_112\bin\javaw.exe (06-Jan-2018, 4:35:41 PM)
RecommendedItem[item:3003, value:3.807882]
RecommendedItem[item:1008, value:3.2369695]
RecommendedItem[item:2003, value:2.9798841]
```

Figure 2: Recommendation for User ID 15

Product number 2008, which has a recommendation rating of 4, is advised that User 11 buy. It is advised that User 11, Bajaj Auto Ltd., take into consideration purchasing nuts from Sundaram Fasteners Ltd. The product's high recommendation value of 4 further supports the model's endorsement of the particular maker. The next two entries on the list are 3003 and 1008, with recommendation values of 3.9 and 3.2, respectively.

Likewise, it is advised that user 15 adhere to item 3003, 1008, and 2003 in order. Based on the algorithm's analysis, it is advised that User #15 think about buying Product #3003, which has a recommendation value of 3.8, should User #15 show signs of being inclined to buy.



```
<terminated> EvaluateRecommender [Java Application] C:\Program Files\Java\jre1.8.0_112\bin\javaw.exe (20-Jan-2018, 7:51:54 PM)
Evaluation value = 0.3699736503454355
Precision = 0.7857142857142857
Recall = 0.7857142857142857
```

Figure 3: Output showing evaluation, precision, and recall

In this investigation, two separate metrics—precision and recall—are employed. We may obtain the accuracy and recall metrics for the recommender system under consideration by using the built-in methods getPrecision() and getRecall(). Over 80% of investigations have shown that the suggested model accurately predicts outcomes, with a precision and recall rate of 0.78. The results of the recommendation engine examination are shown in Figure 6.9. The degree to which precision and recall may be computed is determined by how comparable the two sets of data are.

VI. CONCLUSION AND FUTURE SCOPE

Currently, one of the hardest data mining problems to solve is one that considers customer privacy concerns. Scholars place great trust in big data mining. As a result, a research topic might be selected. The research yields a ground-breaking decision tree method for safeguarding private data. It has been suggested that data with vertical partitions be subjected to data mining. The suggested method is contrasted with the cutting-edge C4.5 algorithm, which operates on unpartitioned data but does not preserve user privacy. The core of the operation is conducted in Weka, which is the industry leading Big Data Mining software. The recommended algorithm outperforms cutting-edge methods in both centralized and decentralized setups, according to the results.

This study also presents a new Big Data Application, a Recommender System, which forecasts the Fasteners market's future need by utilizing Apache Mahout. The suggested approach makes fastener recommendations based on historical purchase trends in the business sector. In turn, this aids manufacturers of fastening products in projecting future demand.

We have put out a novel legal strategy to address the difficulty of safeguarding people's privacy in the Big Data era. The need for a privacy-preserving framework in the Big Data era is made abundantly clear by the evidence around the legislative procedures that have previously been in place to prohibit the misuse of private data. Thus, the suggested paradigm is quite helpful in this sense. Four groups can be formed from the study's results.

REFRENCES:

[1]. Keller, F., Muller, E., & Bohm, K. (2012, April). HiCS: high contrast subspaces for density-based outlier ranking. In Data Engineering (ICDE), 2012 IEEE 28th International Conference on (pp. 1037-1048). IEEE.

[2]. Kriegel, H. P., Kroger, P., Schubert, E., & Zimek, A. (2012, December).Outlier detection in arbitrarily oriented subspaces. In Data Mining (ICDM), 2012 IEEE 12th International Conference on (pp. 379-388).

IEEE

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

[4]. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health information science and systems, 2(1), 3.

[5]. Rebentrost, P., Mohseni, M., & Lloyd, S. (2014). Quantum support vector machine for big data classification. Physical review letters, 113(13), 130503.

[6]. Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In Collaboration Technologies and Systems (CTS), 2013 International Conference on (pp. 42- 47). IEEE

[7]. Saha, B., Shah, H., Seth, S., Vijayaraghavan, G., Murthy, A., & Curino, C. (2015, May). Apache tez: A unifying framework for modeling and building data processing applications. In Proceedings of the 2015 ACM SIGMOD international conference on Management of Data (pp. 1357-1369).ACM.

[8]. Shin, J. E., Jung, B. H., & Lim, D. H. (2015). Big data distributed processing system using RHadoop. Journal of the Korean Data and Information Science Society, 26(5), 1155-1166.

[9]. Zhang, Y., Gao, Q., Gao, L., & Wang, C. (2012). imapreduce: A distributed computing framework for iterative computation. Journal of Grid Computing, 10(1), 47-68.