Novel Privacy Preserving Classification Mining Approach Applied to A City Public Security Big Data Analysis

Zhengqi Zhang\textsuperscript{1*}, Jian Wang\textsuperscript{2}

\textsuperscript{1} Shanwei Vocational and Technical College, Shanwei, 516600, China
\textsuperscript{2} Henan University of Economics and Law, Zhengzhou, 450000, China
*Corresponding author’s e-mail: 20112016@huel.edu.cn

Abstract. With the development of the big data analysis technology, more and more companies are willing to apply this technology. In order to further promote the big data investigation teaching of Sichuan Police College, we accurately grasp the new situation, new tasks and new requirements faced by the big data investigation, and further deepen the solidification of the implementation path. Through on-the-spot visit, discussion and communication, we will focus on understanding the application technology of big data, extend it to various work of public security organs’ attack, prevention and control. Then we want to promote the establishment of practical application systems, and promote the fundamental change of criminal investigation mode through the construction of big data.

1. Introduction
The cloud vision does offer some particularly challenging privacy problems that are unlikely to be sufficiently addressed by today’s best practice [1]. Such issues are due to the fact that the input data for cloud services is uploaded by the user to the cloud, which means that they typically result in users’ data being present in unencrypted form on a machine that the user does not own or control. This poses some inherent privacy challenges [2]. As large amounts of data which are stored in the cloud contain private information and are in non-aggregate format, sharing the data with third-party service providers in the cloud without strict restriction will cause great threat to data privacy.

The problem of privacy-preserving data mining has found considerable attention in recent years because of recent concerns on the privacy of underlying data. In recent years, the issue of privacy protection in classification has been raised. The objective of privacy-preserving data classification is to build accurate classifiers without disclosing private information in the data being mined. The performance of privacy-preserving techniques should be analyzed and compared in terms of both the privacy protection of individual data and the predictive accuracy of the constructed classifiers. Although increased awareness of the privacy issues in the cloud, little work has been done in privacy-preserving data classification based cloud environment. Therefore our research provides some contribution to solve the above problems.

At present, the public security organs of G city are trying to introduce artificial intelligence and deep learning into the systematic attack of the core platform of block data command, so that the system can learn the solidified modules and successful cases independently, link the solidified modules into multi-layer organizations. Then it will form a neural network, form a new module and label system, and filter out the clues of investigation and detection in big data analysis.
Make full use of the huge amount of big data resources accumulated by the public security organs in G City in social security prevention and control and criminal case investigation. The aim of us is to achieve intelligent analysis, early warning and research and judgment, make full use of the advantages brought by scientific and technological progress. We also want to reduce the physical effort of the police in the process of social security and criminal case investigation. For those who have a history of crimes, the relevant data are imported into the system for early warning analysis. This activity can prevent investigation work in advance, collect evidence in time and prevent a large number of basic investigation work from acquiring and analyzing data after the event. It can play an effective role in preventing and solving cases for the investigation work of public security organs and become a powerful assistant for investigators.

2. New Problems Encountered
The increasingly heavy task and the contradiction of limited personnel need to rely on higher technical means and platform to solve.

2.1. Cloud platform big data system analysis
After a large amount of data is provided and collected by various hardware terminals and grass-roots police officers, powerful computers are needed to calculate and analyze the value of the data, especially the correlation and correlation analysis between data. For example, the video image data of key personnel captured by the smart camera, or even the video image data of the same key personnel captured many times, all need the computer to compare and analyze its similarity and analyze the time of capturing the image. Then it can make correlation analysis with other data, such as hotel accommodation information, shopping information and other related data, so as to provide reference for the investigators, and ultimately improve the efficiency of investigation. However, not every public security organ or grassroots unit has the ability to establish its own computer hardware system. It has to rely on the cloud computing system uniformly distributed by large social cloud platform enterprises or the ministry of public security for calculation and analysis.

2.2. Big data training and artificial intelligence system development
After a large amount of data is imported, it is necessary to train the algorithm of the analysis system. According to the actual situation of local investigation, it is necessary to train and analyze the big data, so that the artificial intelligence comparison system can analyze and compare according to the actual situation of local investigation as much as possible. Then it can establish its own analysis model for each region, so as to improve the efficiency of the investigation work. The analysis results are compared and corrected to continuously improve the accuracy of the analysis model, so as to improve the efficiency of the investigation work and bring practical results.

2.3. Training of relevant talents in big data field
In the process of investigation, the use of big data analysis system and artificial intelligence comparison system is inseparable from the high end talents of big data analysis and artificial intelligence analysis model. It is very necessary to introduce and cultivate professional talents of big data analysis and artificial intelligence system analysis. And to train and learn in the public security and investigation posts, to become compound high end talents, or to introduce and cultivate professional talents for big data analysis and artificial intelligence system analysis. And then combined with the work experience and analysis ability of experienced investigators, combined with the establishment of relevant analysis model, not in the practice of investigation evolution of relevant algorithm model, continuous optimization, so that the analysis system can not be separated from the right arm of the investigation staff. At the same time, the application of big data system and the reserve of talents in the investigation work will also provide better talent support for the future intelligent police.
3. Binary Weighted Cosine (BWC) Metric To Measure Similarity

Rawat et.al.[3] proposed BWC similarity measure for measuring similarity across sequences of system calls. They showed the effectiveness of the proposed measure on IDS. They applied k-nn classification algorithm with BWC metric measure to enhance the capability of the classifier. BWC similarity measure considers both the number of shared elements between two sets as well as frequencies of those elements in traces. The similarity measure between two sequences A and B is given by:

$$S(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \cdot \frac{|A \cap B|}{|A \cup B|}$$

Assume A and B are sets containing elements. The Algorithm 1 shows the steps required to calculate the BWC similarity function.

Algorithm 1: BWC similarity function: S(A,B)

```
int S_intersection = 0; S_union = 0;
for all a \in A and b \in B {
    if (a=b) {
        S_intersection = S_intersection + 1;
        S_union = S_union + 1;
    } else
        S_union = S_union + 2;    }
BWC Notation S(A, B);
S(A, B) = \frac{S_intersection}{S_union};
```

Binary Weighted Cosine (BWC) Metric similarity measure is derived from Cosine similarity as well as Jaccard similarity measure. Since the Cosine similarity measure is a contributing component in a BWC similarity measure hence, BWC similarity measure is also a vector based similarity measure. The transformation step is same as carried out in Cosine similarity measure or Euclidean measure for sets. For two sets, A= { p, s, t, n, q, k, r, m } and B={ k, m, t, k, m, q, m } , the computed BWC similarity measure comes out to be 0.45.

4. Privacy Preserving BWC Similarity Function

The following is the algorithm to compute BWC similarity value in privacy preserving manner.

Algorithm 2: Privacy Preserving BWC Similarity Function

```
int S_intersection = 0; S_union = 0;
for all a \in A and b \in B {
    if ( PMP (a, b) = True)                                {
        S_intersection = S_intersection + 1;
        S_union = S_union + 1;                                        }
    else S_union = S_union + 2;    }
BWC Notation S(A, B);
S(A, B) = \frac{S_intersection}{S_union};
```


else \( S_{\text{union}} = S_{\text{union}} + 2; \) 

BWC Notation \( S(A, B); \)

\[
\begin{align*}
S(A, B) &= \frac{S_{\text{intersection}}}{S_{\text{union}}};
\end{align*}
\]

5. K-NN Classification Algorithm For Cloud Data

Privacy-preserving data mining becomes an important and practical technology for mining data from multiple private databases owned by different and possibly competing organizations. For example, many insurance companies collect data on disease incidents, seriousness of the disease and patient background. One way for the Center for Disease Control to identify disease outbreaks is to train a classifier across the data held by the various insurance companies for patterns that are suggestive of disease emergence and use it to classify a query pattern as an emergence or the opposite. However, commercial and legal reasons prevent the insurance companies from revealing their data. It is necessary and beneficial to use a distributed data mining algorithm that is capable of identifying potential disease emergence while protecting the private information of its participants.

Based on the concept of big data computing, the public security organs of G City organize criminal investigation experts who have long been engaged in the fight against crime to sort out all kinds of criminal cases. We get the related information from the law of criminal cases, the characteristics of crime means, the behavior mode and personnel characteristics of criminal crowd. And the practical application of investigation thinking in criminal cases are summarized and refined, and dozens of expert investigation modules, including major criminal case investigation, robbery and snatch case investigation, burglary case investigation, network fraud case investigation, gun trafficking case investigation, etc. are implanted into the core platform. Criminal records, drug abuse records, regional criminal groups, long-term use of rental cars, abnormal capital flow, the use of black cards and other business labels of more than 200 types, are gathered through the establishment of a supporting command and coordination mechanism. It significantly improved the ability of the public security organs in G city to prevent crimes and improve the efficiency of investigating and solving cases.

With the concern and support of party committees and governments at all levels, in recent years, the public security organs have seized the historical opportunity of big data strategic action and police mechanism reform. Through several years of construction and development, the infrastructure and hardware conditions for information construction have taken shape, based on the super data carrier cloud platform, relying on the Skynet video private network covering the whole region, and other information perception networks. Therefore a comprehensive, multi-level, automated three-dimensional information perception network has been built to ensure the integrity, freshness and accuracy of information data. At present and in the future, the main direction of the construction of public security organs' big data search is to continue to consolidate the hardware construction, accelerate the software construction, support the service guarantee, transform and upgrade the practical application, and strengthen the training of practical talents.

In Algorithm 4, the Broker will consider one record of a selected database as a test record. It sends this test record to all the nodes in the cloud. A pair (class label, similarity values) of k neighbour records is gathered in a set called Local Common set and they are sent to the classifier. As a result the classifier will get \( k(n - 1) \) number of entries. Now the classifier will consolidate such lists from all the nodes. It will merge them into a single Global common list. From this list it will pick the k most similar nodes and assign the test record with the class name using the weighted k-NN approach.

Algorithm 3: Local Classification

Consider the records in database a as the training dataset.

Consider the records b as the test record from database B.

Calculate its BWC similarity with respect to all the records of training set through Algorithm 3

Pick the nearest k records from the training set.
Get the records matching the most with the testing record.
Consider these records as the Local Common set for A.
Return a vector containing the pair
(class label, similarity value) of each record for all the k records to the Broker.

Algorithm 4: k-nn Classification

Parties involved: Let the P1,P2,...Pn be the participating training nodes with databases
DB1, DB2,...DBn, where n is the number of parties. Assume a semi-honest third party calles
Broker will pick the k nearest neighbours
at global level.
Local phase: Consider the node P1 as the testing node
for all records in the testing node database
{
   for all training nodes {
      Compute k local neighbour using the
      Algorithm 4.
      Send pairs (class label, similarity value)
      of k local neighbour records to the Broker.
   }
}

Global phase:
Merge all the pairs for a training record as a single list called the global list.
Broker send global list to classifier.
Classifier selects k most similar records based on the similarity measure at the global level
Assign the class label for the test record based on the weights (similarity value) of
the k neighbors.

6. Conclusion
Under the guidance of the core concept of big data, the public security organs of G city construct the
expert model base and label system in the way of human-computer interaction based on the core
platform of block data command. It also combined with the professional data of criminal investigation,
technical investigation, network investigation and video investigation. By using big data online
analysis, stream vector calculation and other operation modes, a data-driven investigation mode of
systematic and large-scale cooperative operation is formed, which is supported by the core platform of
block data command, led by criminal investigation and coordinated by multiple police forces.

References
[1] Sadie Creese, Paul Hopkins, Siani Pearson, Yun Shen, “Data protection-aware design for cloud
computing,” In Proceedings of CloudCom 2009, Beijing, Springer LNCS, December 2009.
[2] Jian Wang, Jiajin Le, “Based on Private Matching and Min-Attribute Generalization for Privacy
Preserving in Cloud Computing,” The sixth International Conference on Intelligent
Information Hiding and Multimedia Signal Processing, IIH-MSP 2010, Darmstadt, Germany,
October, 2010.
[3] Rawat, S. Pujari, A.K., Gulati, V.P., and Vemuri, V. Rao, “Intrusion Detection using Text
Processing Techniques with a Binary-Weighted Cosine Metric,” International Journal of
Information Security, Springer-Verlag, 2004.