Structured Data Watermarking Algorithm based on Deep Stochastic Forest Models

Yali Peng, Deliang Gong, Fang Fang*
Xiangnan University, Chenzhou, 423000, China

*Corresponding author(E-mail: quan_su09@163.com)

Abstract
With the rapid development of the computer communication technology, computer network in the country’s political, national defense, culture, economy, education, health, science and technology, public services, and other fields has been widely used that has brought great convenience to human beings. Traditional network security technology, such as information encryption technology, can encrypt the data and information into unrecognized ciphertext through encryption key and encryption algorithm, which can protect the security of information transmission in public channel. Unauthorized persons who are unable to know the contents of the confidential information, but it will be easy to perceive the existence of confidential information. As a new research field, information hiding technology involves many fields, such as data hiding technology, image processing technology, multimedia technology, cryptography, coding and algorithm design. Under this basis, this paper presents the novel data watermarking algorithm based on deep stochastic forest models. The performance is well verified through the experimental results. Compared with the other models, the proposed one can obtain better performance.

Key words: Structured Data, Watermarking, Deep Stochastic Forest, Data Mining, Information Security.

1. INTRODUCTION

In recent years, information hiding has become a new research direction in the field of the information technology. It has been widely used in copyright protection, secret transmission and storage, hidden channel and anonymous technology. Using information hiding technology, important confidential information can be hidden in the public digital file, by passing public digital files to pass the secret information, so that third parties cannot detect the existence of secret information or even secret communication, so as to achieve safe delivery of the information purpose.

Information hiding technology is the need to keep confidential information, hidden in the non-confidential information while making it in the appearance of the form is only a general content containing information. Information hiding technology in accordance with the treatment of the different objects, divided into image technology, digital watermarking technology and alternative technology. (1) Video-based information hiding technology: The information hiding in video is primarily concerned with the need to be compatible with the internationally common video compression encoding. Hiding algorithm based on the embedded video in time to the information hiding algorithm based on video can be divided into information hiding in front of the video coding and video encoding of information hiding. Before the video code, the hidden information is embedded in the un-encoded raw video and then encoding the video embedded in the secret information. (2) Image-based information hiding technology: As the image has a large redundant space, the image processing tools more and hide the amount of information is better and easy to detect, where the information hiding technology is widely used. (3) Audio-based information hiding technology: Audio information hiding technology is an important branch of information hiding technology, which is currently used mainly in the copyright protection of audio information and the secret transmission of audio information. The difference between the hidden technologies of the audio message is that the data is embedded and extracted in different ways.

At present, information hiding technology is summarized in the five aspects in the various fields of the information security. (1) Copyright protection of digital works: Service providers in grant works to users at the same time, the both sides of the code in the form of the watermark information hidden in the works, the watermark theoretically should not be destroyed. When digital works are found to be illegally propagated, they can be traced by extracting watermarks. (2) Data integrity: Regarding the data integrity confirmation is must confirm in the data on-line transmission or the memory process has not been tampered with. It is tampered with once through the use frail watermark technology protection media can destroy the watermark, while thus is also distinguished very easily. (3) Data privacy: Transmitting some data over the Internet to prevent unauthorized users from intercepting and using it is an important part of network security. (4) Non-repudiation of data: This can use the watermarking technique to send or receive information in any form of the transaction system, adding the respective signatures to the message in the form of a watermark to achieve the purpose of confirming its behaviour. This watermark should not be removed of the main content. (5) Security enhancement: Business
activities and various bills of security is also the use of the information hiding technology. In the digital notes hidden in the watermark after printing still exists, we can re-scan back to the digital form to extract the security watermark and verify the authenticity of the bill.

Figure 1. The Structured Data Watermarking Framework

Unstructured data usually has two significant features that are closely related to data scale growth and data types that are closely related to the application. For example, for the same video type, video-on-demand applications and video search applications are treated differently for this file. For example, many of the video on demand is based on the point of time to carry out a simple positioning, and search may also involve through keywords or samples for content-based search. For unstructured data management systems, scalable distributed storage mechanisms and mechanisms for synthesizing all types of the unstructured data are research hotspots for unstructured management issues. Under this basis, this paper presents the novel structured data watermarking algorithm based on deep stochastic forest models.

2. THE THEORETICAL BASIS

2.1 The Random Forest

Random forest was first proposed by the Leo Breiman of the University of California in 2001. It is the combination classifier composed of many basic classifiers "decision tree". The decision trees are independent and identically distributed. When a test sample is input, the voting result of all the basic classifiers is used to determine the final sample category. The traditional stochastic forest through founds a series of independently classifies the sample with the distributed decision tree, comes the policy-making final classification result with the voting result. The stochastic forest has introduced two randomisation processes that enables the different decision tree sorter to have the different classified ability, some decision tree's classified performance is good, another some decision tree's classified performance is bad, but, in determined when which category attribute a sample does belong to, these two kind of decision trees have the same voting weight, thus can weaken the sorter the overall performance. The figure 2 presents the general topology.

Figure 2. The Random Forest Network Topology
Extreme random forest classifier is superior to random forest classifier in terms of classification accuracy and training time cost, but extreme random forest classifier can only support offline training mode and does not support incremental learning. Therefore, we mainly consider the following aspects.

- Using statistics of incremental learning algorithm are often not stored samples, but through the study of the related data statistics and analysis of the huge amounts of data classifier extension to obtain the necessary information, statistical sample is discarded no longer use, so when dealing with small sample data, unable to provide adequate statistical information and make the incremental learning failure. To this end, we consider using stored samples incremental learning, in the subsequent incremental learning process will still be able to repeat using historical information, so even for small sample data flow can be stable increment learning.
- According to the degree of chaos of the sample set, the growth rate of the tree can be controlled and the sensitivity to the data noise can be reduced.
- Traditional incremental learning method based on single decision tree need to ensure that decision tree all the decision attributes is optimal, and classifier integration methods, such as extreme random forest algorithm, each decision tree all don't need for the optimal decision attribute, so using the integration method can avoid the greed of each decision tree reconstruction. At the same time, single decision tree over fitting problems, whereas the integration approach could effectively avoid these problems, and can reduce the bias and variance but also can keep the calculation efficiency.

![Figure 3. The Random Forest Inner Organizations](image)

As reflected in the figure 3, the general organizations are demonstrated. Due to the different classification ability of the decision tree, and some decision tree classification is effective and others have poor classification effect and the weight of the corresponding decision tree should be set according to the classification ability. In order to improve the robustness of weights, an improved random forest classifier is constructed by secondary training. The results of a training re-introduced into the classifier, while the second training, to strengthen the classification of excellent decision-making tree weight, weaken the poor classification of the decision-making tree weight, to enhance the classification level. Firstly, we define the model as equation 1 and 2.

\[
\begin{align*}
    y &= \|m(\theta)\| \\
    \|y\| &= \sqrt{\sum_{i=1}^{P} y_i^2}
\end{align*}
\]

RF is a combination of decision trees, using the bagging method to generate different training sets, that is, from the original training set using bootstrap sampling to generate a new training set, for each new training set, the use of random feature selection method to generate decision tree, and the decision trees do not prune during growth. And the parameter can be defined as the follows.

\[
\theta_i(y) = \cos^{-1}\left( y_i \sqrt{\prod_{j=0}^{i-1} \sin \theta_j} \right)
\]

Where the \( \prod_{j=0}^{i-1} \sin \theta_j \) presents the potential information area, RF as the simplest random feature selection is to randomly select a group (eg. F) input variables at each node so that the node segmentation of the decision tree is based on the selected features of the F, rather than examining all the features to decide. Then, the tree is completely grown using the CART method, without pruning, helping to reduce tree bias.
As reflected in the formula 4, the performance is well organized, and for the analysis of the dichotomous variables with or without, birth or death, occurrence or non-occurrence, the method of the logistic regression is generally used. Logistic regression is essentially as a distinction between the two variables of the dependent variable Y. Logistic regression is very sensitive to the multivariate collinearity of the independent variables, requiring independent variables to be independent of that each other. The generalization error is used as the evaluation criterion of generalization ability. First determine the optimal parameters on the training set, and then train the model on the entire training set, and finally use the test set to calculate the generalization error, the process iteration 25 times. The figure 4 shows the performance of random forest.

\[
\ln \| y_n \| = \beta_0 + \sum_{k=1}^{K} \beta_k \ln x_{kn} + \sum_{j=1}^{L} \beta_j \ln \theta_j + \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{L} \beta_{kl} \ln x_{kl} \ln x_{kl} \tag{4}
\]

\[\text{Figure 4. The Performance of Random Forest Compared with Other Models}\]

\section{2.2 The Deep Network}

AutoEncoder automatic encoder is the use of the artificial neural network with a hierarchical structure, assuming that the output and input are the same, and then use unlabeled data training to adjust the network parameters of the method. The working principle is: the unlabeled data input to an encoder, get the code, through the decoder code decoding to lose a message, this information and the input signal constitutes an error, by adjusting the encoder and decoder parameters, making the reconstruction error is minimal, encoding code is a representation of the input signal. At present, in the depth of the large data analysis used in neural network mainly feedforward neural network and as the name implies, data flow in the network is one-way: from the first layer (input layer) inflows, passed on (hidden) step by step and mapping, flowing from the last layer (output layer). The depth of the network, namely the number of neurons in the neural network layers, by increasing the number of hidden layer, each layer to relay the characteristics of the original data with the method of learning, is essentially in approaching the strong nonlinear mapping relationship between original data and its characteristics. According to the principle of uniform approximation of the neural network, for an arbitrary nonlinear mapping, must be able to find a shallow network and a deep network approach it with any degree of accuracy, as long as the number of hidden layer neurons enough shallow network or deep web deep enough. But usually, the more shallow network, deep web only need fewer parameters can achieve the same approximation effect.

\[
\ln L = \sum_{n=1}^{N} \left[ -\frac{1}{2} \ln (2\pi) - \ln \sigma_s - \ln \Phi \left( \frac{z_n \delta}{\sqrt{\sigma_s^2 \gamma}} \right) + \ln \Phi \left( \frac{(1-\gamma)z_n \delta - \gamma \varepsilon_i}{\sqrt{\sigma_s^2 \gamma (1-\gamma)}} \right) - \frac{1}{2} \frac{\left( \varepsilon_i + z_n \delta \right)^2}{\sigma_s^2} \right] \tag{5}
\]

The formula 5 expresses the objective function for the proposed model, deep belief network is a kind of containing the deep structure of multiple restricted Boltzmann machine, as a method of feature extraction can be
combined by the low-level features to form a more abstract executive said with the performance. RBM effective learning process makes it suitable as a DBN constituent module, the last layer of RBM output information as BP neural network input data. Thus, DBN can also be considered as a neural network with training initial weights with the following boundary conditions that should be considered.

\[ \left| f'(x) - f'(y) \right| > \delta \]

\[ x_2 = x_2e; \frac{\partial v_1}{\partial x_2} = 0, \frac{\partial v_2}{\partial x_2} = 0, \frac{\partial T}{\partial x_2} = 0, \]

\[ \frac{\partial e_a}{\partial x_2} = 0, \frac{\partial U_R}{\partial x_2} + \frac{c}{2} U_R = 0. \]

Structureally, since the encoder is a forward acyclic network, contains an input layer, a hidden layer, an output layer. A self-encoding network is formed when the encoder contains multiple hidden layers. The number of hidden nodes in the depth from the encoder is usually significantly less than the number of the input nodes, forming a compressed network structure, so the last hidden layer of the activation response can be seen as the compressed representation of the input sample as the follows.

![Figure 5. The Sample Deep Network Simulation Results](image)

3. THE PROPOSED MODEL

3.1 The Structured Data Paradigm

Unstructured data must be corresponding interpretation software to open it and with intuitive browsing, therefore, it cannot be directly obtained from the data itself is the expression of physical properties, which is not easy to understand. Unstructured data, especially the multimedia data is very large amount of information, if stored in the database directly, in the addition to greatly increase the capacity of the database that can reduce maintenance and application efficiency, particularly for the general small and medium-sized database system. Unstructured data does not have the rigid structure, so it is more difficult to standardize and then manage than structured information. In the system, we should contain the following components.

- Query machine cluster, the user in the query machine to issue a query command to establish a query plan, query machine metadata node cluster based on the metadata stored information, to the storage node distribution query task, and finally summarize multiple storage nodes to return the query results submitted to the user with the finalized information.
- Loading machine cluster, the load data of the whole system can be in the process, on multiple devices at the same time set up multiple concurrent load client to improve the system as a whole by concurrent load loading efficiency.
- Storage node cluster, persistent storage of long-term preservation of historical data. The data source is stored in blocks, usually one or more times from the loader to refresh the data in the cluster as a data block unit and the general interactive system of the data structures.
- Metadata node cluster: used to coordinate the whole cluster, metadata information preservation work of the whole system. The required metadata node cluster storage includes a data node status information; index slice storage specific location information; spatial metadata tables; each table in the space of that some auxiliary information and system log etc.
In the log, flow record applications, such as time property is the most commonly used to retrieve attributes as MDSS attribute is used for the data partition management, index fragmentation between holding time orderly. At the same time, based on time property distributed B + Tree, speed up the partition data retrieval process. B + Tree leaf nodes corresponding to the maximum time property and save each index divided index fragmentation of the location of the storage nodes. Distributed B + Tree is well stored in the metadata node in the cluster. The figure 6 demonstrates the general forms of the system. And the data can be separated from listed perspectives. (1) Unstructured database is in view of the relational database model is too simple, the inconvenience to express complex nested needs and support data types such as limited that from the perspective of the data model and put forward comprehensive new database theory based on the Internet application. Unstructured database compared with the currently popular relational database, the biggest difference is that it breaks through the relational database data structure definition is not easy to change and fixed limits, support the repeat field, sub-fields and variable-length field and implements to deal with the variable length data and repeated fields and longer storage management, data items in processing continuous information (including full text information) and unstructured information (repeat) and variable length data has incomparable advantage over traditional relational database. (2) Multimedia database is a database system that can effectively realize the functions of the storing, reading and retrieving multimedia data. Multimedia database is the general product of database technology and multimedia technology and multimedia database from the multimedia data and the characteristics of the information itself, taking into account the introduction of the database after the relevant problems.

Figure 6. The Structured Data Structure and Paradigm Demonstration

3.2 The Data Watermarking Algorithm

Compared with the still image, audio, video, 3D watermark embedding and extraction is more complicated, the spatial information of the model data or the transform domain information is modified to achieve the basic purpose of embedded watermark. When the embedded information is too large, the original carrier data model, it will be distorted as embedded watermark is easy to be detected. For this reason, many scholars have adopted the visual masking technology of the human visual system, which makes the watermark embedding process more complicated and has a long computation time. The three-dimensional model can be divided into two kinds of boundary representation and spatial representation. Polygon modeling is the most widely used modeling technique. It is the most commonly used expression method of the current 3D model. However, in the practical application, the polygon grid for a triangular mesh, the information describing a 3D grid model includes: vertex coordinates, polygons and attribute information, as a grid is the approximation of the shape of an object, so this paper uses a three-dimensional triangular mesh model.

\begin{equation}
P(i \text{ chooses } A) = \frac{\exp(\beta \cdot U_i(A))}{\exp(\beta \cdot U_i(A)) + \exp(\beta \cdot U_i(B))}
\end{equation}

\begin{equation}
P(i \text{ chooses } B) = \frac{\exp(\beta \cdot U_i(B))}{\exp(\beta \cdot U_i(A)) + \exp(\beta \cdot U_i(B))}
\end{equation}

826
As shown in the formula 8 and 9, the original probability for the watermarking system is presented. However, after reviewing the existing models, we can summarize the following drawbacks. (1) Since the watermark embedding algorithm and the detection algorithm are reciprocal, or symmetrical, the watermarking algorithm cannot be made public. (2) Because watermark’s embedding and watermarking are all used in the same password, the required public secret key must be disclosed in the extraction process.

It is decided by the existence of these two problems, only by the copyright owner or watermark will be embedded watermark existence discrimination, this limits the practical application of the watermarking algorithm, and planted the seeds of the above two problems exist, because has the character of symmetric and public key algorithm itself, so that the one can solve the problem of the above two asymmetrical public watermarking algorithm, has the very high application value.

![Figure 7. The Data Watermarking Sample Framework](image)

4. EXPERIMENT

The simulation tests mainly demonstrate the invisibility and robustness of watermarks, which are somewhat contradictory in one way or another. In this paper, PCA is used to analyze the principal component of the image. The extracted coefficients are the main components that represent the most important features of the image. These principal components contain both the high frequency part of the image and the low frequency part of the image. The general frequency domain transform DCT, DWT and so are the image of the low frequency domain and high frequency domain strictly separated, and for different attacks, the watermark embedded in a separate frequency domain, the algorithm will show different robustness, such as DWT transform. The geometric attack is less robust. Because the principal component has the characteristics of high frequency and low frequency non-separation, embedding the watermark into the main component extracted by PCA can effectively avoid the problem of general frequency domain transform algorithm. In the following figures, we present the experimental results. The figure 8 shows the watermarking framework of the proposed model. The figure 9 demonstrates the multiple testing of the proposed model. The figure 10 demonstrates the overall testing of the detecting accuracy. It can be reflected that the proposed model performs better than the other approaches.
Figure 8. The watermarking framework of the proposed model

Figure 9. The multiple testing of the proposed model
Figure 10. The overall testing of the detecting accuracy

5. CONCLUSION

Information hiding technology is the need to keep confidential information, hidden in the non-confidential information while making it in the appearance of the form is only a general content containing information. Information hiding technology in accordance with the treatment of the different objects, divided into image technology, digital watermarking technology and alternative technology. This paper presents the novel idea of the information watermarking model, the experimental results show that compared with the traditional DWT watermarking algorithm while the algorithm not only improves the anti-geometric attack ability, such as anti-clipping and rotation, but also shows strong robustness to the attack of noise and image gray value changes. The algorithm can be used in digital content authentication, copyright protection and other aspects of promotion. In the future, we will apply the proposed model into more scenarios to verify the effectiveness.

ACKNOWLEDGEMENT
This paper is financially supported by the Hunan Provincial Department of Education research projects (No.17C1486); Hunan Science and Technology Department of Science and Technology Project (2015GK2051); 2015 Chenzhou City Science and Technology Bureau general projects.

REFERENCES


Gubanov, M., Stonebraker, M., & Bruckner, D. (2014). “Text and structured data fusion in data tamer at scale”. In Data Engineering (ICDE), 2014 IEEE 30th International Conference on (pp. 1258-1261). IEEE.


