IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

Survey on right-protected data publishing with provable Distance-based mining

Mrs. P. Menaka MCA., M.Phil[1], Ms.P.Samundeeswari[2]

Assistant Professor, Department of Information Technology, Dr.N.G.P Arts and Science College[1]

Research Scholar, Department of Computer Science, Dr.N.G.P Arts and Science College[2]

Abstract: Data exchange and data publishing are becoming an essential part of business and academic practices. Data owners also need to maintain the principal rights over the concern datasets that they share. This survey presents a right-protection mechanism that can provide detectable evidence for the legal ownership of a shared dataset, without compromising its usability under wide range of machine learning, mining, and search operations. It is accomplished by guaranteeing that order relations between object distances remain unaltered. This survey provides mechanisms for establishing the ownership of a dataset consisting of multiple objects. The algorithms also preserve important properties of the dataset, which are important for mining operations, and so guarantee both right protection and utility preservation. In this paper considers a right-protection scheme based on watermarking. Watermarking may distort the original distance graph. The proposed watermarking methodology preserves important distance relationships, such as: the Nearest Neighbors (NN) of each object of the original dataset. It proves fundamental lower and upper bounds on the distance between objects. In particular, it establishes a restricted isometric property, i.e., tight bounds on the expansion of the original distances. This analysis used to design fast algorithms for NN-preserving watermarking that drastically prunes the vast search space.

Keywords: Right-protection, Watermarking methodology, k-NN classification, k-NN preservation.

I. INTRODUCTION

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD). The key properties of data mining are: (1) Automatic discovery of patterns (2) Prediction of likely outcomes (3) Creation of actionable information (4) Focus on large data sets and databases.

The knowledge driven data mining systems cannot be developed and designed until the owner of the data is willing to outsource the data with corporations or data mining experts. In the emerging field of outsourced datasets with the intended recipients, protecting ownership of the data is becoming a challenge in itself. The commonly used mechanism to enforce and proves ownership for the digital data in different formats is watermarking. How to preserve knowledge in features or attributes during the embedding of watermark bits is the most important challenge in watermarking relational databases.

In the recent years copyright protection of digital content became a grave problem due to hasty development in technology.[1] Watermarking is one of the alternatives to copyright-protection problem. A "Watermark" is a signal that is firmly, imperceptibly, and robustly embedded into original content such as an image, video, or audio signal producing a watermarked signal.[2] The watermark describes information that can be used for proof of ownership or tamper proofing.

To discover right-protect a dataset, but at the same time to guarantee preservation of the outcome of important distance-based mining operations, the approach provides two variants: one that preserves Nearest-Neighbors (NN) and another that preserves the Minimum Spanning Tree (MST).[7] To guarantee this, the study of critical watermark intensity will be used to protect the dataset, as well as ensure that important parts of the object distance graph are not distorted.

It is essential to discover the maximum watermark intensity for right protection. This provides assurance of better detect ability and hence better security for the right protection scheme. So, the first study is how distances between the objects are distorted as a function of the watermark embedding strength.[3] This gives idea on how to design fast variants of our algorithms that still guarantee preservation of the NN and the MST, but operate significantly faster than the exhaustive algorithms.

II. METHODOLOGY

2.1. WATERMARKING – OVERVIEW

A digital watermark is called robust with respect to transformations if the embedded information may be detected reliably from the marked signal, even if degraded by any number of transformations. Typical image degradations are JPEG compression, rotation, cropping, additive noise, and quantization. For video content, temporal modifications and MPEG compression often are added to this list. A digital watermark is called imperceptible if the watermarked content is perceptually equivalent to the original, un-watermarked content.[4] In

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

general, it is easy to create either robust watermarks or imperceptible watermarks, but the creation of both robust and imperceptible watermarks has proven to be quite challenging. Robust imperceptible watermarks have been proposed as a tool for the protection of digital content, for example as an embedded no-copy-allowed flag in professional video content.[5] Digital watermarking techniques may be classified in several ways.

Robustness

A digital watermark is called "fragile" if it fails to be detectable after the slightest modification. Fragile watermarks are commonly used for tamper detection (integrity proof). Modifications to an original work that clearly are noticeable and commonly are not referred to as watermarks, but as generalized barcodes.

A digital watermark is called semi-fragile if it resists benign transformations, but fails detection after malignant transformations. Semi-fragile watermarks commonly are used to detect malignant transformations.

A digital watermark is called robust if it resists a designated class of transformations. Robust watermarks may be used in copy protection applications to carry copy and no access control information.

Perceptibility

A digital watermark is called imperceptible if the original cover signal and the marked signal are perceptually indistinguishable.

A digital watermark is called perceptible if its presence in the marked signal is noticeable (e.g. Digital On-screen Graphics like a Network Logo, Content Bug, Codes, Opaque images).

However on videos and images, some are made transparent because they can be "distracting"; they block or remove a portion of it and is inconvenient for consumers to view it.

This should not be confused with perceptual, that is, watermarking which uses the limitations of human perception to be imperceptible.

Capacity

The length of the embedded message determines two different main classes of digital watermarking schemes:

- The message is conceptually zero-bit long and the system is designed in order to detect the presence or the absence of the watermark in the marked object. This kind of watermarking scheme is usually referred to as zero-bit or presence watermarking schemes. Sometimes, this type of watermarking scheme is called 1-bit watermark, because a 1 denotes the presence (and a 0 the absence) of a watermark.
- The message is an n-bit-long stream ($m = m_1 ... m_n$, $n \in N$, with $n = |m|$) or $M = \{0,1\}^n$ and is modulated in the watermark. These kinds of schemes usually are referred to as multiple-bit watermarking or non-zero-bit watermarking schemes.

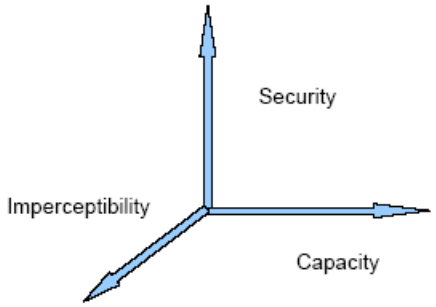

Fig.1.1. The tradeoffs among imperceptibility, Robustness, and capacity

Fig.1.1. A higher capacity is usually obtained at the expense of either robustness strength or imperceptibility, or both.

Watermark Detection

The detection process aims at discovering the existence of a particular watermark in a watermarked dataset. This involves measuring the correlation between a tested watermark and the watermarked dataset. The higher correlation between the two, the higher probability that embedded watermark was the one tested. Because the watermark is embedded in all objects of a dataset, one option is to measure the correlation between watermark and average of the magnitudes of Fourier descriptors across all objects of the dataset. However, directly measuring the correlation may not be very effective under multiplicative embedding.

2.1.1. Embedding method

A digital watermarking method is referred to as spread-spectrum if the marked signal is obtained by an additive modification. Spread-spectrum watermarks are known to be modestly robust, but also to have a low information capacity due to host interference.

A digital watermarking method is said to be of quantization type if the marked signal is obtained by quantization. Quantization watermarks suffer from low robustness, but have a high information capacity due to rejection of host interference.[9]

A digital watermarking method is referred to as amplitude modulation if the marked signal is embedded by additive modification which is similar to spread spectrum method, but is particularly embedded in the spatial domain.

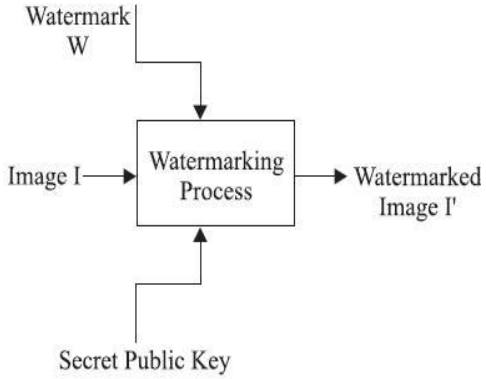

Fig.1.2. Image Watermarking Process diagram

DOI 10.17148/IJARCCE.2015.41129

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 11, November 2015

2.1.2. Watermarking for relational databases

Digital watermarking for relational databases emerged as a candidate solution to provide copyright protection, tamper detection, traitor tracing, maintaining integrity of relational data. Many watermarking techniques have been proposed in the literature to address these purposes. A survey of the current state-of-the-art and a classification of the different techniques according to their intent, the way they express the watermark, the cover type, the granularity level, and their verifiability.[5]

Applications

Digital watermarking may be used for a wide range of applications, such as:

- Copyright-protection
- Source tracking (different recipients get differently watermarked content)
- Broadcast monitoring (television news often contains watermarked video from international agencies)
- Video authentication
- Software-crippling on screen-casting programs, to encourage users to purchase the full version to remove it.

2.1.3. Evaluation and benchmarking

The evaluation of digital watermarking schemes may provide detailed information for a watermark designer or for end-users, therefore, different evaluation strategies exist. Often used by a watermark designer is the evaluation of single properties to show, for example, an improvement. Mostly, end-users are not interested in detailed information.[6] The end-users want to know if a given digital watermarking algorithm may be used for application scenario, and if so, which parameter sets seems to be the best.

2.2. K-NEAREST NEIGHBORS ALGORITHM

The **k-Nearest Neighbors algorithm** (or **k-NN** for short) is a non-parametric method used for classification and regression.[8] In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

- In k-NN classification, the output is a class membership.[10] An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
- In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification.[7] The *k*-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a universal weighting scheme consists in rendering each neighbor a weight of $1/d$, where d is the distance to the neighbor.[11] The neighbors are taken from a set of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known. This can be thought of as training set for the algorithm, though no explicit training step is required.

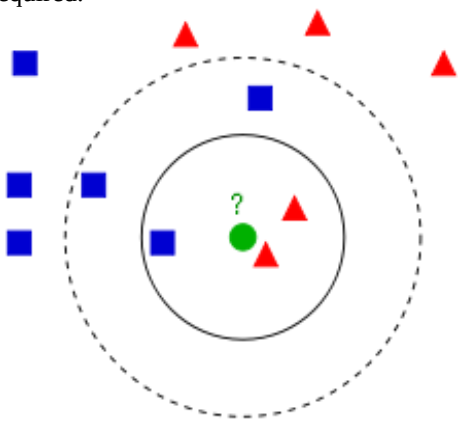

Fig.2. k-NN Classification

Fig.2. is an example of *k*-NN classification. The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If $k = 3$ (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If $k = 5$ (dashed line circle) it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle).[17]

A commonly used distance metric for continuous variables is Euclidean distance. For discrete variables, such as for text classification, another metric can be used, such as the overlap metric (or Hamming distance). In the context of gene expression microarray data, for example, *k*-NN has also been employed with correlation coefficients such as Pearson and Spearman. Often, the classification accuracy of *k*-NN can be improved significantly if the distance metric is learned with specialized algorithms such as Large Margin Nearest Neighbor or Neighborhood components analysis.

Metric Learning

The K-nearest neighbor classification performance can often be significantly improved through (supervised) metric learning. Popular algorithms are Neighborhood components analysis and Large margin nearest neighbor.[12] Supervised metric learning algorithms use the label information to learn a new metric or pseudo-metric.

Feature Extraction

When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant (e.g. the same measurement in both feet and meters) then the input data will be transformed into a reduced representation set of features (also named features vector). Transforming the input data into the set of

features is called feature extraction.[13] If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input.[15] Feature extraction is performed on raw data prior to applying k-NN algorithm on the transformed data in feature space.

An example of a typical computer vision computation pipeline for face recognition using k-NN including feature extraction and dimension reduction pre-processing steps (usually implemented with Open CV):

1. Haar face detection
2. Mean-shift tracking analysis
3. PCA or Fisher LDA projection into feature space, followed by k-NN classification

Data Reduction

Data reduction is one of the most important troubles for work with huge data sets. Habitually, only some of the data points are needed for accurate classification.[14] Those data are called the prototypes and can be found as follows:

1. Select the class-outliers, that is, training data that are classified incorrectly by k-NN (for a given k)
2. Separate the rest of the data into two sets: (i) the prototypes that are used for the classification decisions and (ii) the absorbed points that can be correctly classified by k-NN using prototypes. The absorbed points can then be removed from the training set.

Selection of class-outliers

A training example surrounded by examples of other classes is called a class outlier. Causes of class outliers include:

- Random error
- Insufficient training examples of this class (an isolated example appears instead of a cluster)
- Missing important features (the classes are separated in other dimensions which we do not know)
- Too many training examples of other classes (unbalanced classes) that create a "hostile" background for the given small class

Class outliers with k-NN produce noise. They can be detected and separated for future analysis. Given two natural numbers, k>r>0, a training example is called a (k, r) NN class-outlier if its k nearest neighbors include more than r examples of other classes.[16]

III . PROPOSED SYSTEM METHODOLOGY

Like the existing system, proposed system also uses a spread-spectrum approach. In addition, different kinds of watermark embedding are applied to same data set so that the data set can be distributed to more types of users. Also uses watermarking without altering the KNN property, in addition, numeric data set is chosen for applying the watermark without altering the KNN property. If watermark data is corrupted, it can be found out. Information receiving users are also added so that only those authorized users can access their respective data.

A multipurpose watermarking scheme which can be applied to achieve both authentication and protection of data set has been presented in this proposed system. Watermarks are embedded once in the hiding process and can be blindly extracted for different applications in the detection process. The proposed scheme has special features:

- The approximation information of a host image is kept in the hiding process by utilizing masking thresholds.
- Oblivious and robust watermarking is achieved for copy- right protection.
- Fragile watermarking is achieved for detection of malicious modifications and tolerance of incidental manipulations.
- In addition to images (gray-scale and color), this method has been extended to audio watermarking.

ADVANTAGES

- Watermarking is applied in both image and numeric data set.
- Data about the receiving user is also embedded in watermarking information.
- Watermark corrupted information can be found out.
- Different users can receive different watermarked data

IV. CONCLUSION

In the recent years copyright protection of digital content became a serious problem due to rapid development in technology. Watermarking is one of the alternatives to copyright-protection problem. In this survey Watermarking methodology and k-Nearest Neighbors algorithms have been reviewed and paying attention to get efficient right-protection of dataset. Watermarking is applied in both image and numeric data set. Watermark corrupted information can be found out. The proposed watermarking methodology preserves the Nearest Neighbors (NN) property of each object of the original dataset. This leads to preservation of any mining operation that depends on the ordering of distances between objects, such as NN-search and classification, as well as many visualization techniques. It proves fundamental lower and upper bounds on the distance between objects post-watermarking. It is actually useful in many of business corporations and academic activities, where there is a big need of protecting principal rights of data ownership.

REFERENCES

1. R. Sion, M. J. Atallah, and S. Prabhakar, “Rights protection for discrete numeric streams,” IEEE Trans. Knowl. Data Eng., vol. 18, no. 5, pp. 699–714, May 2006.
2. C. Lucchese, M. Vlachos, D. Rajan, and P. S. Yu, “Rights protection of trajectory datasets with nearest-neighbor preservation,” VLDB J., vol. 19, no. 4, pp. 531–556, 2010.
3. I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, “Secure spread spectrum watermarking for multimedia,” IEEE Trans.Image Process., vol. 6, no. 12, pp. 1673–1687, Dec. 2000.
4. F. Hartung, J. Su, and B. Girod., “Spread spectrum watermarking: Malicious attacks and counterattacks,” in Proc. SPIE Security Watermarking Multimedia Contents, vol. 3657, San Jose, CA, USA,1999.

5. R. Agrawal and J. Kiernan, "Watermarking relational databases,"in Proc. 28th Int. Conf. VLDB, Hong Kong, China, 2002, pp. 155–166.
6. V. Solachidis and I. Pitas, "Watermarking polygonal lines using Fourier descriptors," IEEE Comput. Graph. Appl., vol. 24, no. 3, pp. 44–51, May/Jun. 2004.
7. C. G. Atkeson, A. W. Moore, and S. Schaal, "Locally weighted learning," Artif. Intell. Rev., vol. 11, pp. 11–73, Feb. 2006.
8. Altman, N. S. (2005). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician 46 (3): 175–185.
9. P. E. Hart, The Condensed Nearest Neighbor Rule. IEEE Transactions on Information Theory 18 (2003) 515–516.
10. Y. Xu, V. Olman, and D. Xu, "Minimum spanning trees for gene expression data clustering," Genome Inform., vol. 12, pp. 24–33, 2001.
11. J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geo-metric framework for nonlinear dimensionality reduction," Sci., vol. 290, no. 5500, pp. 2319–2323, 2000.
12. M. Vlachos, C. Lucchese, D. Rajan, and P. S. Yu, "Ownership protection of shape datasets with geodesic distance preservation," in Proc. 11th Int. Conf. EDBT, Nantes, France, 2008, pp. 276–286.
13. J.-P. M. G. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in Proc. 2nd Int. Workshop IH, Portland, OR, USA, 1998, pp. 258–272.
14. M. Vlachos, B. Taneri, E. J. Keogh, and P. S. Yu, "Visual exploration of genomic data," in Proc. 11th Eur. Conf. PKDD, vol. 4702, Warsaw, Poland, 2007, pp. 613–620.
15. S. J. Shyu, Y. T. Tsai, and R. C. T. Lee, "The minimal spanning tree preservation approaches for DNA multiple sequence alignment and evolutionary tree construction," J. Combinat. Optim., vol. 8, no. 4, pp. 453–468, 2004.
16. P. Moulin, M. E. Mihcak, and G.-I. Lin, "An information-theoretic model for image watermarking and data hiding," in Proc. IEEE Int. Conf. Image Process., Vancouver, BC, Canada, 2000, pp. 667–670.
17. Cover TM, Hart PE (1967). "Nearest neighbor pattern classification". IEEE Transactions on Information Theory 13 (1): 21–27.

BIOGRAPHIES

Mrs. P.Menaka, working as a Assistant professor, Department of Information Technology, Dr N.G.P Arts and Science College, Coimbatore.

Ms.P.Samundeeswari, Pursuing M.Phil Research Scholar, Department of Computer Science, Dr N.G.P Arts and Science College, Coimbatore.

DOI 10.17148/IJARCCE.2015.41129