|| Volume 2||Issue 8 ||FEB 2017||ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

Privacy Preserving Data Mining Using Piecewise Vector Quantization

Suhas Uttamrao Shinde
Master of Engineering ,Computer Science and Engineering Department,
Everest Educational Society's College of Engineering and Technology, Aurangabad, Maharashtra, India

***Abstract*– Most content sharing websites allow users to enter their privacy preferences. Unfortunately, recent studies have shown that users struggle to set up and maintain such privacy settings. In this paper, we propose an Adaptive Privacy Policy Prediction (A3P) system which aims to provide users a hassle free privacy settings experience by automatically generating personalized policies. The A3P system handles user uploaded images, and factors in the following criteria that influence one's privacy settings of images. We design the interaction flows between the two building blocks to balance the benefits from meeting personal characteristics and obtaining community advice.**

Keywords- PPDM, Vector Quantization, A3P, Data Mining, APP, Prediction.

I INTRODUCTION

Data mining is a technique that deals with the extraction of hidden predictive information from large database. It uses sophisticated algorithms for the process of sorting through large amounts of data sets and picking out relevant information. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. With the amount of data doubling each year, more data is gathered and data mining is becoming an increasingly important tool to transform this data into information. Long process of research and product development evolved data mining. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.

Most of the techniques for privacy preserving data mining (PPDM) uses modified version of standard data mining algorithms, where the modifications usually using well known cryptographic techniques ensure the required privacy for the application for which the technique was designed. In most cases, the constraints for PPDM are preserving accuracy of the data and the generated models and the performance of the mining process while maintaining the privacy constraints. The several approaches used by PPDM can be summarized as below: [2]

1. The data is altered before delivering it to the data miner.
2. The data is distributed between two or more sites, which cooperate using a semi-honest protocol to learn global data mining results without revealing any information about the data at their individual sites.
3. While using a model to classify data, the classification results are only revealed to the designated party, who does not learn anything else other that the classification results, but can check for presence of certain rules without revealing the rules.

The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information. A number of techniques such as randomization and k-anonymity have been suggested in recent years in order to perform privacy-preserving data mining. Furthermore, the problem has been discussed in multiple communities such as the database community, the statistical disclosure control community and the cryptography community. In some cases, the different communities have explored parallel lines of work which are quite similar.[3]

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks.

Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes. We tried to preserve privacy of data by applying VQ on segments of the datasets and transforming datasets into new datasets. This method segmentizes every point (row) of datasets into w segments each of length L. Each segment is represented by the closest, based on

distance measure, entry in the codebook(containing code vector of L dimension).The point (row) of data, which is segmentizes, is now represented by new transformed row of data formed by joining the new segments formed. In order to create the codebook, each point (row) data in a training set is partitioned into w segments, each of length l. These segments form the samples used to train the codebook using K means clustering method. Codebook contains the centroid of all clusters formed through K means clustering method. Each segment of length L is approximated and represented by centroid, of length L, of cluster in which it falls as resulted by K means clustering method i.e. closest based on distance measure in codebook .These new w segments formed are joined to form a new point (row) which is replacement of original point (row) of dataset to which this w segments belong. Thus preserving privacy as dataset is completely transformed to new dataset. It also gives accuracy in terms of clusters as similarities between the points in datasets are preserved. Similarity depends on Euclidean distance between points which is more or less preserved.

II LITERATURE SURVEY

Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

III SYSTEM ARCHITECTURE

System Construction Module

The A3P system consists of two main components: A3P-core and A3P-social. The overall data flow is the following. When a user uploads an image, the image will be first sent to the A3P-core. The A3P-core classifies the image and determines whether there is a need to invoke the A3P-social. In most cases, the A3P-core predicts policies for the users directly based on their historical behavior. If one of the following two cases is verified true, A3P-core will invoke A3Psocial: (i) The user does not have enough data for the type of the uploaded image to conduct policy prediction; (ii) The A3P-core detects the recent major changes among the user's community about their privacy practices along with user's increase of social networking activities (addition of new friends, new posts on one's profile etc.). The A3P system handles user uploaded images, and factors in the following criteria that influence one's privacy settings of images. We design the interaction flows between the two building blocks to balance the benefits from meeting personal characteristics and obtaining community advice. Data is mined to anticipate behavior patterns and trends.

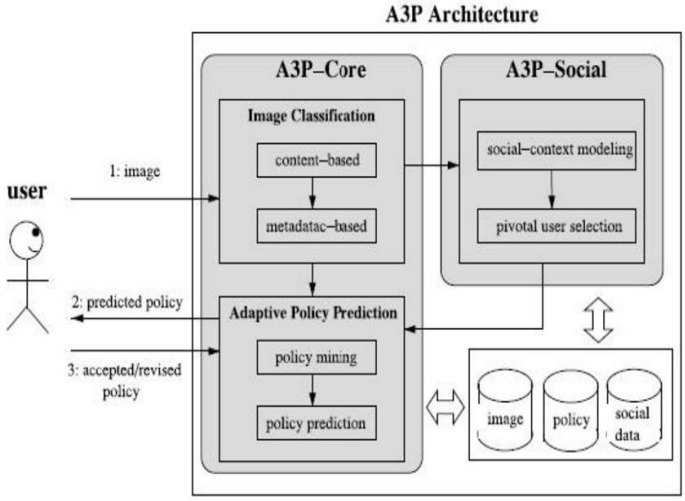

Figure 1: The A3P Architecture

Content-Based Classification

To obtain groups of images that may be associated with similar privacy preferences, we propose a hierarchical image classification which classifies images first based on their contents and then refine each category into subcategories based on their metadata. Images that do not have metadata will be grouped only by content. Such a hierarchical classification gives a higher priority to image content and minimizes the influence of missing tags. Note that it is possible that some images are included in multiple categories as long as they contain the typical content features or metadata of those categories. Our approach to content-based classification is based on an efficient and yet accurate image similarity approach. Specifically, our classification algorithm compares image signatures defined based on quantified and sanitized version of Haar wavelet transformation. For each image, the wavelet transform encodes frequency and spatial information related to image color, size, invariant transform, shape, texture, symmetry, etc. Then, a small number of coefficients are selected to form the signature of the image. The content similarity among images is then determined by the distance among their image signatures.

Metadata-Based Classification

The metadata-based classification groups images into subcategories under aforementioned baseline categories. The process consists of three main steps. The first step is to extract keywords from the metadata associated with an image. The metadata considered in our work are tags, captions, and comments. The second step is to derive a representative hypernym (denoted as h) from each metadata vector. The third step is to find a subcategory that an image belongs to. This is an incremental procedure. At the beginning, the first image forms a subcategory as itself and the representative hypernyms of the image becomes the subcategory's representative hypernyms.

Adaptive Policy Prediction

The policy prediction algorithm provides a predicted policy of a newly uploaded image to the user for his/her reference. More importantly, the predicted policy will reflect the possible changes

of a user's privacy concerns. The prediction process consists of three main phases: (i) policy normalization; (ii) policy mining; and (iii) policy prediction.

IV CONCLUSION

This project gives a different approach of using piecewise vector quantization for privacy preserving clustering in data mining. In this, we have showed analytically and experimentally that Privacy-Preserving Clustering is to some extent possible using piecewise vector quantization approach. Using piecewise quantization approach a data owner can meet privacy requirements without much losing the benefit of clustering since the similarity between data points is preserved or marginally changed.

V FUTURE SCOPE

As future work new and effective quantization method can be used rather than K means approach that we have used. K nearest neighbor approach is one of the approach which can give better result in more work in the field of fuzzy Dataset, Mobility of different Dataset, The development of uniform framework for various privacy preserving across all data mining algorithms.

REFERENCES

[1] IS. Sasikala, IIS. NathiraBanu, "Privacy Preserving Data Mining Using Piecewise Vector Quantization (PVQ)", IJARCST 2014, VOL. 2, Issue 3 (July – Sept. 2014).

[2] Boyang "Oruta: Privacy-Preserving Public Auditing for Shared Data in the Cloud", IEEE TRANSACTIONS ON CLOUD COMPUTING, VOL. 2, NO. 1, JANUARY-MARCH 2014.

[3] LidanShou, He Bai, Ke Chen, and Gang Chen "Supporting Privacy Protection in Personalized Web Search", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 2, FEBRUARY 2014.

[4] D.ArunaKumari, Dr.K.Rajasekharrao, M.suman "Privacy preserving distributed data mining using steganography "In Procc. Of CNSA-2010, Springer Library

[5] T.Anuradha, sumanM, ArunaKumari D "Data obscuration in privacy preserving data mining in Procc International conference on web sciences ICWS 2009.

[6] Ciriani V, Vimercati SDC, Foresti S, Samarati P (2008) k-anonymous data mining: a survey. In: Privacy-preserving data mining. Springer, New York, pp 105–136

[7] Aggarwal CC, Yu PS (2008) a general survey of privacy-preserving data mining models and algorithms. In: Privacy preserving data mining, Chap 2. Springer, New York, pp 11–52

[8] Arunadevi M, Anuradha R (2014) Privacy preserving outsourcing for frequent itemset mining. Int J Innov Res Comp Commun Eng 2(1):3867–3873

[9] Chan J, Keng J (2013) Privacy protection in outsourced association rule mining using distributed servers and its privacy notions, pp 1–5

[10] Deivanai P, Nayahi J, Kavitha V (2011) a hybrid data anonymization integrated with suppression for preserving privacy in mining multi party data. In: IEEE international conference on recent trends in information technology (ICRTIT)