INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A SURVEY ON: DIFFERENT TECHNIQUES AND FEATURES OF DATA CLASSIFICATION

[1]Manisha Kannvdiya, [2]Kailash Patidar, [3]Rishi Singh Kushwaha

Sri Satya Sai Institute of Science and Technology, Sehore (SSSIST)
Pachama Sehore, Madhya Pradesh, India- 466001

Abstract: - As the number of internet users are increasing day by day. This increase in number was done because of the different available services, with dynamic functionality. So this attracts various researchers for finding interesting and fruitful field for research. One of the major fields is privacy and security of the individual data. This paper has focus on the classification service provide by the servers with privacy of the uploaded data of the client. Paper has explained techniques with various features use for classification. Here paper has explained different evaluation methods for the technique performance checking as well.

Keywords:-Image processing, Support Vector machine Text classification, ANN.

I. INTRODUCTION

As the data miners are gathering information from the large dataset base on useful patterns, trends, etc. This is useful for helping crime understanding, any kind of terrorist activity can also be learn by the data mining approach. Classification between the objects is easy task for humans but it has proved to be a complex problem for machines. The raise of high-capacity computers, the availability of high quality and low-priced video cameras, and the increasing need for automatic video analysis has generated an interest in object classification algorithms. A simple classification system consists of a camera fixed high above the interested zone, where images are captured and consequently processed.

Mining [1] is the discovery by computer of new previously unknown information by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar within web search. In search user is looking for

something already known and has been written by someone else. The problem is pushing aside all material that currently is not relevant to your needs in order to find relevant information. In text mining, the goal is to discover unknown information, something that no one yet known and so could not have yet written down. It is an interdisciplinary field involving information retrieval.

Privacy of data which contain some kind of medical information about the individual, financial information of family or any class. As this make some changes on the dataset, so present information in the dataset get modify and make it general for all class or rearrange so that miner not reach to concern person.

II. D ATA CLASSIFICATION TECHNIQUES

KNN (K Nearest Neighbors algorithm) in [4] is used which utilize nearest neighbor property among the items. This algorithm is easy to implement with high validity and required no prior training parameters. Although K nearest neighbor is also identified as instance based learning in other words classification of items is quite slow. In this classification techniques distance between the K cluster center and classifying item is calculated then assign item to cluster having minimum distance from the cluster center. In case of text mining features from the document is extracted then k labeled node is select randomly which are supposed to be cluster center and rest of nodes or document are unlabeled nodes. Finally distance between labeled and unlabeled node is calculate on the base of feature vector similarity. In this algorithm distance between nodes are estimate in log(k) time .

Advantages: Main significance of this algorithm is that this is robust against raw data which contain noise. In this algorithm prior training is not required as done in most of the neural network for classification. One more flexibility of this algorithm is that this work well in two or multiclass partition.

Limitations: In this work selection of appropriate neighbor is quite high if population of item is large in number. One more issue is that it required much time for finding the similarity between the document features. Because of these limitations this algorithm is not practical with large number of items. So cost of classification increases with increase in number of items.

Support Vector Machine (SVM) in [3] is quite famous soft computing technique for item classification which is based on the input feature vector quality and training of the support vector machine. In this technique an hyperplane is built between the items this hyperplane classify the items into binary or multi class. In order to find the hyperplane equation is written as P = B+XxW where X ia an item to be classify then W is vector while B is constant. Here W and B is obtained by the training of SVM. So SVM can perfectly classify binary items by using that calculated hyperplane.

Advantages: Main significance of the Support Vector Machines is that it is less susceptible for over fitting of the feature input from the input items, this is because SVM is independent of feature space. Here classification accuracy with SVM is quite impressive or high. SVM is fast accurate while training as well as during testing.

Limitations: In this classification multiclass items are not perfectly classify as number of items reduce gap of hyperplane.

Image Classification

Fuzzy classification in [15], has classify image data which is highly complex and required stochastic relations for the creation of feature vector from images. Here different types of relations are combined where members of the

feature vector is fuzzy in nature. So this relation based image classification is highly depending on the type of image format as well as on the threshold selection.

Advantages: This algorithm is easy to handle, while stochastic relation help in identifying the different uncertainty properties.

Limitation: Here deep study is required to develop those stochastic relations; accuracy is depending on prior knowledge.

III. R ELATED WORK

Yu et al. in [5] has proposed a scheme where client can freely provide its data to the un-trusted server for data analysis. Here both type of data such as scalable or fine grained data is analyzed by utilizing the data feature attribute with key policy attribute encryption algorithm KPABE. Here in order to identify the data features are supplied to the server in encrypted manner where feature so extract that server could not regenerate the data even after cracking the encryption algorithm. Here information files are encrypted by random key at client end. Now some of the authorized users can classify the data that have correct decryption key. But in this work classification owner required number of different keys for sending data on the server. So making number of group and updating of those keys on regular interval is highly required.

Lu et al. in [6] has proposed a provenance approach where ownership of records are maintain and process history of information objects. Here grouping of owner signature with the chipper test policy with attribute based encryption was done. This chipper policy of attribute based encryption is term as CPABE. By the use of this scheme an authentication is required for the owner to access its files. While pattern of the user was store in the cloud for the user behavior learning as it will alarm for the intruder attack. So by using the ABE any user can encrypt the data file and store it on the server. For accessing the file user need correct signature as input. Here main drawback of the work is that revocation of the personal keys of the data owner is required but it is not done.

Efficient Revocation in CP-ABE Based Cryptographic Cloud Storage. Yong CHENG [7] proposed a security for customers to store and shares their sensitive data in the cryptographic cloud storage. It provides a basic encryption and decryption for providing the security and data confidentiality. However, the cryptographic cloud storage still has some shortcomings in its performance. Firstly, it is inefficient for data owner to distribute the symmetric keys one by one, especially when there are a large number of files shared online. Secondly, the access policy revocation is expensive, because data owner has to retrieve the data, and re-encrypt and re-publish it. The first problem can be resolved by using cipher text policy attribute-based encryption (CP-ABE) algorithm. To optimize the revocation procedure, they present a new efficient revocation scheme. In this scheme, the original

Shobha D. Patil et al, data are first divided into a number of slices, and then published to the cloud storage. When a revocation occurs, the data owner needs only to retrieve one slice, and re-encrypt and re-publish it. Thus, the revocation process is affected by only one slice instead of the whole data.

B. Wang et al. in [8] has proposed a cloud computing algorithm with storage services. Here one more feature of the work is that with storage facility on the cloud work can distribute that data to multiple parties for sharing of information. Here a KNOX approach is proposed which provide privacy preserving for storing and sharing of the information on the cloud. Here a third party is use to verify the access of the files on the cloud for the same it required that proposed work use signature based homomorphic authentication. Although data owner has the power

to add or delete any user to access the data files as per situation or requirement. In this work time required for finding the authentication of the user with amount of required information is quite high.

Dan Boneh in [9], constructs a short group signature scheme with length under 200 bytes where the signatures are nearly the standard RSA signature size with the same level of security. Group signature security of this proposed scheme is based on the Strong Diffie-Hellman (SDH) assumption and a new assumption in bilinear groups called the Decision Linear assumption. This system stands on a new Zero-Knowledge Proof of Knowledge (ZKPK) of the solution to an SDH problem where ZKPK is converted to a group signature via the Fiat-Shamir heuristic.

Fiat et al. in [10] has proposed a approach which efficiently reduces the requirement of the transmission length as well as the storage at the client end. This work includes new theoretical measure for the analysis of the quantitative and qualitative approach. Here an encryption scheme is broadcast which helps in broadcasting transmission. By the use of this approach all group members can efficiently got there respected data files. But this broadcasting has one limitation that group member has not got proper privilege that what kind of information one can read and transfer. So an unauthorized user can also got the file if it in group.

J. Fully Collusion Secure Dynamic Broadcast Encryption with Constant-Size Cipher texts or Decryption Keys. In [11] C. Delerablee introduces new efficient constructions for public-key broadcast which offer stateless receivers, collusion-secure encryption, and high security. In the standard model; new users can join anytime without implying modification of user decryption keys or permanently revoke any group of users. This system achieve the optimal bound of O(1)-size either for cipher texts or decryption keys, also provides a dynamic broadcast encryption system improving all previous efficiency measures (for both execution time and sizes) in the private key setting.

IV. F EATURES FOR CLASSIFICATION

Text Feature

1) Title feature: The word in sentence that also occurs in title gives high score. This is determined by counting the number of matches between the content word in a sentence and word in the title. In [4] calculate the score for this feature which is the ratio of number of words in the sentence that occur in the title over the number of words in the title.

2) Sentence Length: This feature is useful to filter out short sentence such as datelines and author names commonly found in the news articles the short sentences are not expected to belong to the summary. In [5] use the length of sentence, which is the ratio of the number of words occurring in the sentence over the words occurring in the longest sentence of the documents.

3) Term Weight: The frequency of the term occurrence with documents has been used for calculating the importance of sentence. The score of a sentence can be calculated as the sum of the score of words the sentences. The score of important score wi of word i can be calculated by traditional tf.idf method.

Image Features:

1) Color feature: Image is a matrix of light intensity values; these intensity values represent different kind of color. So to identify an object colure is an important feature, one important property of this feature is low computation cost. Different Image files available in different color formats like images have different colure format ranging from RGB which stand for red, green, and blue. This is a three dimensional representation of a single image in which two dimensional matrix represent single color and collection of those matrix tends to third dimension. In order to make

intensity calculation for each pixel gray format is use, which is a two dimension values range from 0 to 255. In case of binary format which is a black and white color matrix whose values are only 0 or 1.

2) Edge Feature: As image is a collection of intensity values, and with the sudden change in the values of an image one important feature arises as the Edge as shown in figure 4. This feature is use for different type of image object detection such as building on a scene, roads, etc. [7]. There are many algorithm has been developed to effectively point out all the images of the image or frames which are Sobel, perwitt, canny, etc. out of these algorithms canny edge detection is one of the best algorithm to find all possible boundaries of an images.

3) Corner Feature: In order to stabilize the video frames in case of moving camera it require the difference between the two frames which are point out by the corner feature in the image or frame. So by finding the corner position of the two frames one can detect resize the window in original view. This feature is also use to find the angles as well as the distance between the object of the two different frames. As they represent point in the image so it is use to track the target object.

V. CONCLUSION

As main goal of the data miners is to retrieve information from the raw or arrange data. From the different common approach for classification of user text, image or data files, supervised classification approach is highly famous. This survey paper has contributed the theoretical explanation of various approaches followed or proposed by different researchers. Paper has given brief introduction of features for the different type of data. So a algorithm is still need to develop for the reduced time and space complexity without compromising classification accuracy.

VI. REFERENCES

[1] Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, "Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases" In IEEE Systems Journal, VOL. 7, NO. 3, SEPTEMBER 2013, pp. 385-395.

[2] C. Tai, P. S. Yu, and M. Chen, "K-support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining," in Proc. Int. Knowledge Discovery Data Mining, 2010, pp. 473–482.

[3] Christoph Goller, Joachim Löning, Thilo Will and Werner Wolff, 2009, "Automatic Document Classification: A thorough Evaluation of various Methods", "doi=10.1.1.90.966".

[4] B S Harish, D S Guru and S Manjunath, 2010, "Representation and Classification of Text Documents: A Brief Review", IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition" RTIPPR..

[5] S. Yu, C. Wang, K. Ren, And W. Lou, "Achieving Secure, Scalable, And Fine-Grained Data Access Control In Cloud Computing," Proc. IEEE INFOCOM, Pp. 534-542, 2010.

[6] R. Lu, X. Lin, X. Liang, And X. Shen, "Secure Provenance: The Essential Of Bread And Butter Of Data Forensics In Cloud Computing," Proc. ACM Symp. Information, Computer and Comm.Security, Pp. 282-292, 2010.

[7] Yong CHENG, Jun MA and Zhi-Ying "Efficient Revocation In Cipertext-Policy Attribute-Based Encryption Based Cryptographic Cloud Storage" Zhejiang University And Springer-Verlag Berlin 2013.

[8] B. Wang, B. Li, And H. Li, "Knox: Privacy-Preserving Auditing For Shared Data With Large Groups In The Cloud," Proc. 10th Int'l Conf. Applied Cryptography And Network Security, Pp. 507-525, 2012

[9] D. Boneh, X. Boyen, And H. Shacham, "Short Group Signature," Proc. Int'l Cryptology Conf. Advances In Cryptology (CRYPTO), Pp.41-55, 2004.

[10] A. Fiat And M. Naor, "Broadcast Encryption," Proc. Int'l Cryptology Conf. Advances In Cryptology (CRYPTO), Pp. 480-491, 1993.

[11] C. Delerablee, P. Paillier, And D. Pointcheval, "Fully Collusion Secure Dynamic Broadcast Encryption with Constant-Size Cipher texts or Decryption Keys," Proc. First Int'l Conf. Pairing-Based Cryptography, Pp. 39-59, 2007.

[12] Sara Hajian and Josep Domingo-Ferrer. "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining". IEEE transaction on knowledge data engineering, VOL. 25, NO. 7, JULY 2013.

[13] Mohamed R. Fouad, Khaled Elbassioni, and Elisa Bertino. A Super modularity-Based Differential Privacy Preserving Algorithm for Data Anonymization. IEEE transaction on knowledge data engineering VOL. 26, NO. 7, JULY 2014

[14] Yogachandran Rahulamathavan, Raphael C.-W. Phan, Suresh Veluru, Kanapathippillai Cumanan and Muttukrishnan Rajarajan."Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud ".IEEE transaction on dependable and secure computing, VOL. 11, NO. 5, September 2014.

[15] Sabna Sharma, Pratikshya Sharma. "Comparative Study on Supervised and Unsupervised Fuzzy Approach for Image Classification". International Journal of Engineering Research & Technology (IJERT). Vol. 3 Issue 5, May - 2014