International Journal of Current Trends in Engineering & Technology
www.ijctet.org, ISSN: 2395-3152
Volume: 04, Issue: 03 (May- June, 2018)

A Survey on Various Classification Techniques for Different Type of Data

Afshan Idrees, Avinash Sharma
Computer Science & Engineering Department
Millennium Institute of Technology, Bhopal India
afshanidrees@yahoo.in
8602299179

Abstract— *As the quantity of web clients are expanding each day. This expansion in number was done in view of the diverse accessible administrations, with dynamic usefulness. So this draw in different scientists for finding intriguing and productive field for inquires about. One of the real fields is protection and security of the individual information. This paper has concentrated on the classification developed by the servers for the security of the transferred information of the customer. Paper has clarified strategies with different elements use for classification. Here paper has clarified distinctive assessment parameters for the system evaluation.*

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

I. INTRODUCTION

As the information mineworkers are gathering data from the vast dataset base on valuable examples, patterns, and so on. This is valuable for helping crime seeing, any sort of fear based oppressor movement can likewise be learning by the information mining approach. Classification between the items is simple task for people however it has turned out to be an intricate issue for machines. The raise of high-limit computers, the accessibility of high caliber and low-valued camcorders, and the expanding requirement for programmed video investigation has produced an enthusiasm for protest order calculations. A basic order framework comprises of a camera settled high over the intrigued zone, where pictures are caught and thus processed.

Mining [1] is the revelation by computer of new already obscure data via naturally removing data from various composed assets. A key component is the connecting together of the extracted data together to shape new certainties or new speculations to be investigated facilitate by more ordinary methods for experimentation. Content mining is not the same as what are well-known inside web seek. In scan client is searching for something definitely known and has been composed by another person. The issue is pushing aside all material that right now is not pertinent to your requirements so as to discover significant data. In content mining, the objective is to find unknown data, something that nobody yet known thus couldn't have yet recorded. It is an interdisciplinary field including data recovery. Security of information which contain some sort of medicinal data about the individual, monetary data of family or any class as this roll out a few improvements on the dataset, so present data in the dataset get adjust and make it general for all class or revamp so miner not reach to concern individual.

II. DATA CLASSIFICATION TECHNIQUES

KNN (K Nearest Neighbors calculation) in [4] is utilized which use closest neighbor property among the things. This calculation is anything but difficult to actualize with high legitimacy and required no earlier preparing parameters. In spite of the fact that K closest neighbor is additionally recognized as occasion based learning as it were order of things is very moderate. In this classification procedures separate between the K bunch focus and characterizing thing is figured at that point allocate thing to group having least separation from the bunch focus. If there should arise an occurrence of content mining highlights from the archive is extricated then k named hub is select haphazardly which are assume to be group focus and rest of hubs or record are unlabeled hubs. At last separation amongst named and unlabeled hub is figure on the base of highlight vector similitude. In this calculation separate between hubs are assess in log (k) time.

Favorable circumstances: Main criticalness of this calculation is this is strong against crude information which contain clamor. In this calculation earlier preparing is not required as done in the greater part of the neural system for characterization. One greater adaptability of this calculation is that this function admirably in two or multiclass parcel.

Restrictions: In this work determination of fitting neighbor is very high if populace of thing is substantial in number. One more issue is that it required much time for finding the comparability between the report highlights. In view of these confinements this calculation is not pragmatic with huge number of things. So cost of arrangement increments with increment in number of things.

Bolster Vector Machine (SVM) in [3] is very well known delicate registering method for thing arrangement which depends on the info highlight vector quality and preparing of the help vector machine. In this system a hyper plane is work between the things this hyper plane characterize the things into paired or multi class. Keeping in mind the end goal to discover the hyper plane condition is composed as P = B+XxW. Where X is a thing to be group then W is vector while B is steady. Here W and B are acquired by the preparation of SVM. So SVM

can impeccably characterize parallel things by utilizing that figured hyper plane.

Favorable circumstances: Main importance of the Support Vector Machines is that it is less powerless for over fitting of the component contribution from the info things, this is on the grounds that SVM is autonomous of highlight space. Here classification precision with SVM is very amazing or high. SVM is quick precise while preparing and also amid testing.

Impediments: - In this order multiclass things are not consummately arrange as number of things lessen hole of hyper plane.

Picture Classification: - Fluffy arrangement in [15] has order picture information which is very intricate and required stochastic relations for the production of highlight vector from pictures. Here various sorts of relations are consolidated where individuals from the element vector is fluffy in nature. So this connection based picture arrangement is exceptionally rely upon the sort of picture design and on the limit determination.

Focal points: - This calculation is anything but difficult to deal with, while stochastic connection helps in distinguishing the diverse uncertainty properties.

Confinement: - Here profound investigation is required to build up that stochastic connection, exactness is relying upon earlier learning.

III. RELATED 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 is 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 an 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. FEATURES FOR CLASSIFICATION

4.1 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.

4.2 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 an 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 Issues 5, May – 2014.