IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, February 2015.
www.ijiset.com
ISSN 2348 – 7968

Online Feature Selection and Stability Analysis Using Data Mining

Bhamidi Sai Karthik[1], and Mr. N. Praveen[2]

[1] Post Graduate Scholar, Department of Computer Science & Engineering, SRM University, Chennai, Tamilnadu, India

[2]Assistant Professor, Department of Computer Science & Engineering, SRM University, Chennai, Tamilnadu, India

Abstract

Using data mining and its techniques with the rapid developments, there has been a growing trend to use the mining services for large-scale data storage. This has raised the issue of a security which is necessary for using how to control and prevent unauthorized access to data stored in the database. Online feature selection is one of the most important mechanisms used in data mining. By using the online learning and transfer learning, features which are used to store the data in web based technology is enhanced through the fine grained access control policies.. Authorized authentication is used to improve security. Transfer learning is provided with authentication that supports write privilege on outsourced data in data mining concept. Classifying this authorization can be provided by specifying the data user's privileges and data owner's policy in data mining techniques. It provides integrity and scalability to the data storage systems efficiently and users will be accessing the data through online.

Keywords: *Wallace Feature Selection, Online Transfer Learning, Data Mining in large data set, Classification*

1. Introduction

Data mining is the delivery of an analytic process through web based resources which include everything from applications to data based technology over the Internet on a pay-for-use basis. Data mining is the particular rule associated for analyzing as well as predicting the user or a customer and stored as a static data in data warehouse to reveal some similar patterns which follows the trend. Mainly these type of techniques are followed by exploration of data which involves the cleaning of data by transforming into data sets consists of large variables known as "FIELDS" with the evolution and adoption of existing technologies and paradigms.

The goal of data mining technique is to allow users to take benefit from all of these technologies, without the necessary about deep knowledge about how to expertise with each one of them. The full input aims to cut costs, and help the users focus on their delivered output instead of being impeded by IT obstacles. Data mining is so named because the information being accessed is found in the extraction of data from a data warehouse, and does not require a user to be in a specific place to gain access to it.

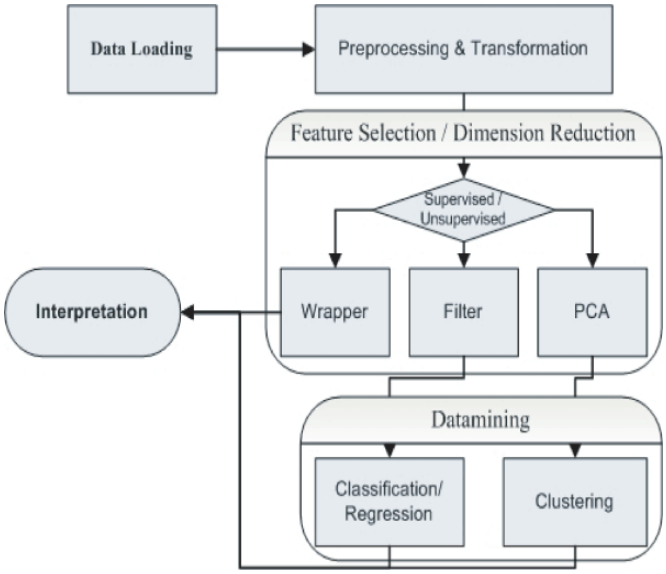

Fig. architecture for proposed system

1.1. INPUT DESIGN

The input design is the link between the 0information system and the user. It comprises the user accessible modulations and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input which has been used for authentication could be focused for controlling the amount of required input, reducing the unwanted errors, avoiding delay with such extra steps and keeping the process simple. The input is designed in

ISSN 2348 – 7968

such a way so that it provides security and ease of use with retaining the privacy.

1.2. OBJECTIVES

Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is created for reducing the errors in the data input process and show the correct direction to the management forgetting. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The aim of resolving a certain input is to make data entry easier and to be free from errors. The data which is created to be designed in such a way that all the data manipulates can be per providing record. When the data is entered it will check for its validity. Data could be validated with the help of presentations produced. Similar messages are to be produced as when needed so that the user will not be in maize of instant. Thus the goal which is abstracted from input design is to create an input layout that is easy to follow.

1.3 EXISTING SYSTEM

Online learning is generally a policy or a procedure that allows the user who get accessed to a system. It also monitors and records all attempts made to access a system. Online learning may also identify users attempting to make an unauthorized access to a system. It is a mechanism which is very much important for providing the details to the user. Various access control models are in use, including the most common Full input Learning, supervised learning, unsupervised learning and Semi supervised learning. All these models are known as identity based access control models. In all these access control models, user (subjects) and resources (objects) are identified by unique names. Identification may be done directly or through authentication assigned to the subjects. These access control methods are effective in unchangeable distributed system, where there are only a set of Users with a known set of services.

Supervised Learning techniques are based on the assumption that the server is in the trusted domain of the data owner and therefore an omniscient reference monitor can be used to enforce access policies against authenticated (Partial) users. However, in the data mining paradigm this assumption usually does not hold and therefore these solutions are not applicable. Therefore, there is a need for a full supervised input which is, scalable and flexible way to control access to web based data without fully relying on the administrator service providers.

Feature Selection has found the applications in many domains especially for the problems involved in high dimensional data. By removing the irrelevant and irredundant features, feature selection can improve the performance of prediction models by alleviating the effect of curse of dimensionality, enhancing the performance of generalization, speeding up the learning process and improving the interoperability model.

1.4 ALGORITHM

Existing Perceptron by Truncation for OFS

1: **Input**

- B: the number of attributes or features

2: **Initialization**

- $w1 = 0$

3: **for** $t = 1, 2. . . T$ **do**

4: Receive x_t

5: Make prediction $\text{sgn}(x_t^\top w_t)$

6: Receive y_t

7: **if** $y_t x_t^\top w_t \leq 0$ **then**

8: $b_{w_{t+1}} = w_t + y_t x_t$

9: $w_{t+1} = \text{Truncate}(b_{w_{t+1}}, B)$

10: **else**

11: $w_{t+1} = w_t$

12: **end if**

13: **end for**

ISSN 2348 – 7968

1.5 DISADVANTAGES OF EXISTING ALGORITHM

- It does not guarantee that numerical values for unselected attributes could be small.
- It decreases the performance when full inputs are selected all at a time.
- User should be aware of the online applications(Internet)
- High Computational Truncation

2. RELATED WORK

Jialei Wang, Peilin Zhao, and Steven C.H.Hoi has proposed a method for implementing a public –key with full input whose data rest in a part on the difficulty of factoring the large number of users, it permits only online learning to be established without the use of fixed number of features on several public data sets.

The Wrapper scheme is able to deal with role hierarchies, whereby role gets or transformed into permissions to form other roles. A user could be capable for joining a role after the owner has encrypted the data for that role. The user will be able to access the full data from then on, and the owner does not need to re-encrypt the data. A user who is registered gets accessed at any time in which case, the revoked user will not have access to any future encrypted data for this role. With the new FS scheme, revocation of a user from a role does not affect other users and roles in the system.

In addition, part of the decryption computation in the scheme to the data base, in which only public parameters are involved is outsourced. By using this approach, the Wrapper scheme achieves an efficient decryption on the client side. The same strategy of outsourcing is used to improve the efficiency of the management of user to role memberships involving only public parameters. OFS provides only coarse grained access control.

By learning with partial input, user cannot just acquire the similar attributes which have non-zero values. This is because, in this way the classifier will never be able to change the subset of attributes with non zero elements for such input data provided.

Instead of encrypting to individual users, in FS, one can embed an access policy into the cipher text or decryption key. Besides, partial input also has collusion-resistance property, i.e., if multiple users collude, they should only be able to decrypt a cipher text if at least one of the users could decrypt it on their own. Thus, data access is self-enforcing from the classification or clustering technique, requiring no trusted mediator. Full input can be viewed as an extension of the notion of identity-based approach in which user identity is generalized to a set of descriptive attributes instead of a single string specifying the user identity.

Compared with unsupervised and semi supervised, Supervised feature has significant advantage as it achieves flexible one-to-many encryption instead of one-to-one, it is envisioned as a promising tool for addressing the problem of partial input as the user can view the full text with non zero elements produced in it and user-defined data gets share to other users with key generated to them by using a high range variety of graphical and statistical methods.
The users can evaluate the performance of the proposed OFS algorithms on large sale data sets which contain at least 100,000 sets and also we can show how the online predictive performance of different algorithms varies over the iteration. Moreover, it is commonly believed that multimodal systems also improve the security against spoofing attacks, which consists of claiming a false identity and submitting at least one fake here and also two types that is filtering and wrapping method then the small amount of data which is been understandable by the user will accepts that biometric trait to the system.

The reason is that to evade multimodal system, one expects the adversary should spoof all the corresponding biometric traits. With this the feature selection explains how the OFS proposed algorithm runs even more faster as the embedded methods rely on the data which gets wrapped into a particular data set on a large scale. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data into a usable form for processing could be achieved by keying the data directly into the system. Therefore, the emphasis is that feature selection should take a decentralized approach to have many users with decrypted key.

To ensure the user authorization, Online Transfer Learning (OTL) should be used. It provides the authentication without disclosing the identity of the users. Other users or the administrator can verify the user and validity of the message stored. It also used to view the data through online so that the full input through which large amount of web based technology could be easily seen when the administrator has its authentication through the user of particular transforming the services. Moreover, the user who log in into the system may be divided into full

ISSN 2348 – 7968

input or partial input due to the administrator does not provide the encryption key to such type of user who was not available in his data in regression technique, when attributes satisfy the access policy.

3. PROBLEM DEFINITION

By evaluating the performance of the proposed OFS algorithms on large data sets which contain throughout some million attributes. The statistics of these data sets are shown in the experiments why the proposed OFS algorithm can even run faster than the other baselines on some data sets. All these desired output results again validate the efficacy and potential of the proposed OFS method for mining large-scale data sets in the era of big data.

4. PROPOSED METHOD

In the proposed method, we are using output design of how the full input data gets accessed to generate the common modulus algorithms by truncation approach which is used to generate the new data.

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Similar and exploration output design could improve the system's relationship to help user decision-making.

1. Designing computer output should proceed in a formed, well similar ascending manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis has been designed to obtain output, the user can identifies the similar output which is needed to meet the requirements.
2. Select methods for presenting information.
3. Create document, report, or other formats that contain information produced by the system. The output obtained from an adequate system should accomplish one or more of the objectives.

Considering partially trustworthy users, to fully realize the data owner centric concept, data owners shall have complete control of their own privacy through encrypting their data files using the OTL based concept under the access of an administrator. The framework addresses the unique challenges brought by multiple data owners and users, greatly reduces the complexity of key management while enhances the privacy guarantees compared with previous works. It authenticates the users who store the information in the database. Truncation and Sparse Projection are used to encrypt the data files, so that data owners can allow access to only authorized and authenticated users.

This method realizes the scalable and fine-grained authentication control in data warehouse under web based applications. By using online multi-class classification as well as regression techniques they can be used as the state of art batch feature selection technique. The following algorithms show how the proposed system makes much faster than the existing system.

Algorithm 1 Modified Greedy by Projection for OFS

w = Truncate (bw, B)

1: **if** k bwk0 > B **then**

2: w = bwB where bwB is bw with everything but the B largest elements set to zero.

3: **else**

4: w = bw

5: **end if**

Algorithm 2 Learning with Partial Inputs. (OTLP)

1: Input

• R: maximum L2 norm

• B: the number of selected features

• ϱ: the enhancement-distribution tradeoff

• η: step size

2: Initialization

• w1 = 0

3: for t = 1, 2. . . T do

4: Sample Zt from an online greedy distribution with randomized ϱ.

5: if Zt = 1 then

6: Randomly choose B attributes Ct from [d]

7: else

8: Choose the attributes that have non-zero values

in wt, i.e., Ct = {i : [wt]i 6= 0}

9: end if

ISSN 2348 – 7968

10: Receive text by only requiring the specified text elements in Ct

11: Make prediction sgn (wTt ext)

12: Receive yt

13: if ytwTt ext ≤ 1 then

14: Compute bxt

15: ewt+1 = wt + ytηbxt

16: bwt+1 = min{1, Rkew t+1k2 } ewt+1

17: wt+1 = Truncate(bwt+1,B)

18: else

19: wt+1 = wt

20: end if

21: end for

4.1 Advantages of Proposed Method

- By combining both the online learning & transfer learning into Single System, its Efficiency could get increased.
- Algorithm depends on system performance with inputs but not by its resources
- Sourcing input makes more efficiency than partial input which does not reoccur

Tab 1 Comparison between existing system and proposed system

S.No	Existing System	Proposed System
1	Only the user can access the application when he had permission.	By Transfer learning all users can access the application through online.
2	Homogeneous constraint network has been implemented.	Heterogeneous constraint network was also produced.
3	Local constraint is not implemented with regression.	Global constraint along with clustering implementation has been proposed.

5. CONCLUSION

By using online multi-class classification as well as regression techniques they can be used as the state of art batch feature selection technique. This could be extended and get proposed into two types of settings. (i). OFS with Full input using OTL method. (ii). OFS with partial input using Regression and OTL method**.** An Online as well as Transfer learning should be provided combining so that the user could be able to learn through full input that was having an encryption key under the particular information. The desired output should be typed in singled-line spacing at the bottom of the page and column where it is cited. Footnotes should be rare.

References

[1] P. Zhao and S. C. H. Hoi. OTL: A framework of online transfer learning. In *ICML*, pages 1231–1238, 2010

[2] P. Zhao, S. C. H. Hoi, R. Jin, and T. Yang. Online auc maximization. In *ICML*, pages 233–240, 2011

[3] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y.Winter. Distributional word clusters vs. words for text categorization. Journal of Machine Learning Research, 3:1183–1208, 2003

[4] J. Bi, K. P. Bennett,M. J. Embrechts, C. M. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. Journal of Machine Learning Research, 3:1229–1243, 2003.

[5] N. Cesa-Bianchi, S. Shalev-Shwartz, and O. Shamir. Efficient learning with partially observed attributes. Journal of Machine Learning Research, pages 2857–2878, 2011

[6] A. B. Chan, N. Vasconcelos, and G. R. G. Lanckriet. Direct convex relaxations of sparse svm. In ICML, pages 145–153, 2007.

[7] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. J. Mach. Learn. Res. (JMLR), 7:551–585, 2006