International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 8, August - 2014

Class Detection of Data Stream using Concept-Drift and Feature-Evolution Techniques for Banking Applications

G. Ruth Rajitha Rani
Vasavi College of Engineering,
Hyderabad

***Abstract*—The rapid advance of computer technologies in data processing, collection, and storage hasprovided unparalleled opportunities to expand capabilities in production, services, communications, and research. However, immense quantities of high-dimensional data renew thechallenges to the state-of-the-art data mining techniques.Analysis of Large (Big) Data and Data mining has become an important study in the field of E-commerce. Data Stream Classification may pose many challenges in the area of Data Mining community. In this paper, an attempt has been made to address the Banking Application Security and Feasibility in Modification of theapplication using Feature (Fe) Evolution and Concept Drift. A general framework for mining concept-drifting data streams using weighted ensemble classifiers is used for this study. A model is trained using these weighted ensemble classifiers.This paper deals with classification technique for Feature-Evolving Data Stream, thereby helping the users to construct more secure information system.**

***Index Terms*—Concept drift, Data Stream, Feature evolution, Novel class.**

I. INTRODUCTION

Security is becoming a critical part of organizational information systems. Feature-Evolving Data Stream is an important detection that is used as a counter measure to preserve data integrity & system availability from intruder attacks.Feature-Evolving Data Stream detects a novel class when a user attempts to bypass the security mechanism of a computer system. Such an attacker can be an outsider who attempts to access the system or an insider who attempts to gain and misuse non-authorized privileges. Feature-Evolving Data Stream is a process of gathering application usage related knowledge in order to monitoring the events & analyzing them for sign or intrusion. It raises an alert when an abnormal class is foundin the system.Based on analysis of data we can evolve feature to train the machine learning algorithm.Misuse detection is based on extensive knowledge of patterns. Class detection is based on a profile that represents normal behavior of users and detecting attacks based on significant deviation from the profiles.

Classification Model is a rule-based.The paper deals with classification by feature evolution. The technique consists of an incremental learning algorithm. In this approach, whenever a new request arrivesbelonging to class c, at first it is checked for usage and then it is added to a vocabulary. After adding, the vocabulary is scanned and statistics are updated. Based on the updated statistics, a novel class is detected.

Based on Session frequency and session time parameters, the decision of whether a new classification model is to be created can be made. Each model in the ensemble is evaluated periodically & old, obsolete models are discarded often.

The above task is done in two ways:

1. Detecting unauthorized user who is trying to access an application.
2. Detecting a user who is using an application for abnormal time period.

II. BACKGROUND

Data Mining tool performs data analysis and may uncover important data pattern, contributing greatly to business strategies & Knowledgebase.In Data Mining where intelligent methods are applied in order to extract data patterns. Data Mining is the process of discovering interesting Knowledge from large amount of data stored either in database, data warehouses, or other information repositories.

The discovered knowledge can be applied to decision making, Process control and information management.

Data Mining tasks can be classified into two categories:

- Descriptive- Characterize the general Properties of the data in the database.
- Predictive- Perform inference on the current data in order to make predictions.

The performance of Data Mining depends on the

- Huge amount of data in the database.
- Data mining algorithm must be efficient & Scalable.

The Dynamic nature of data stream requires efficient and effective techniques. The data stream classification has widely been studied and techniques like Concept-drift, Concept-evolution, and Feature -evolution have been proposed.

In concept-drift technique for novel class detection, an ensemble of model is used to classify the unlabeled data. The novel class detection process consists of three steps:

1. A decision boundary is built during training.
2. Test points falling outside the decision boundary are declared as outliers.
3. The outliers areanalyzed to see if there is enough cohesion among themselves and separation from the existing class instances.

Significant characteristics of data stream are concept-evolution and feature-evolution. In concept-evolution occurs when new classes evolve in the data. Where as in feature-evolution, new features emerge and old features fade away.

The feature –evolution approach maintains a vocabulary. After receiving anunlabeled documentand it is classified and then, vocabulary is updated. Based on new statistics, a new model is created with top N features in its feature vector.

The algorithm decides, based on Session frequency and Session time parameters whether a new classification model is to be created based on parameters periodically and obsolete models are discarded often.

III. METHODOLOGY:

Classification is the process of finding a set of models or functions that describe and distinguish Data classes or Concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data that is data objects whose class label is known.

1.Detection of unauthorized user who is trying to access an application:

When the user is trying to access the application the user has to first login. If the username or password is correct, user will be given access. The user is given three trails to access the application with correct username and password. If the user fails in all three attempts, then the administrator is alerted with the message saying that, someone is trying to intrude.The Database is mined for username and password. The user is authenticated by exact pattern matching with username and password of Database. And even when the frequency of arrival of Username and Password data stream differs after allowing the slack of time, the Database Administrator is alerted. This Concept-drift would recognize a novel class which is intruder class thereby identifying the intruder.

2. Detecting the pattern of usage of application in order to enhance :

This module detects a novel class by the feature-evolution and concept-drift.When the user access the application, his pattern of usage is stored in the database as an ensemble of model. When an request arrivesfor access of the application then a model is constructed with top N features of feature vector. The constructed model is evaluated based on some parameters whether a new classification model is to be created. As models are built, the old fade away. Then the classifier would operate and classify the class label as outlier if there is concept drift.

The usage pattern information is stored as access time and access frequency. When frequency of access or access time after allowing the slack does not match with the stored pattern then there is concept drift and an alert is sent to the administrator.The frequency of access and access time is estimated based in the top two features of session table. The session table would contain username, session time and session date. When data is pooled in to session table of a particular class, the frequency of access and access time is estimated as average of previous frequencies and top two features that is recent and recent past and a slack is allowed for updating the frequencies. When a frequency of usage-pattern would not match with the model and the classifier would classify it as concept-drift and on alert would be sent to administrator.

The figure 1 below shows that when the user is trying to login, the user is given three trails to login .If user fails in these three trails, then the user is shown blank page, and the entry of IP address is recorded in the intruder table and an alert, in the form of a mail is sent to the administrator and the user has to open the application login page for re-login. While logging in, the arrival of packet time is noted and a deviation of 7 is allowed. If the username and password are correct and the frequency of data stream arrival is according to the stored pattern, the user is given access to the application. The pattern ofkeying in of username and password is stored at the time of registration based on the arrival time of packets. Whenever the user logs in, username,password and the frequency of that data stream is verified and then, the user is given access to application.

Fig 2.shows that, after the user logs in, session date is recorded. If the user is logging in for the first time, the number of days frequency of login is recorded as one. That will be the initial pattern (a feature evolved).If the user is a regular user then recent and recent past records are considered to calculatethe number of daysfrequency.The number of days frequency is updated with average of day frequencyand the difference of session dates of recent and recent past (Feature-evolution).Whenever a user logs in,check for concept drift is done by comparing the session date with sum of dayfrequency and previous login date allowing a slack of 7. In case of unusual day frequency a mail is sent to administrator. A thought has to be given for enhancement of application protocol.

Fig 3.Session frequency is current session duration if the record count of session table is one for a particular user.

Session frequency is updated if the difference of average session duration of recent and recent past and Session frequency is less than 7 or else an alert is sent to administrator. A thought has to be given why the user is spending unusual amount of time and there by taking an action to enhance application.

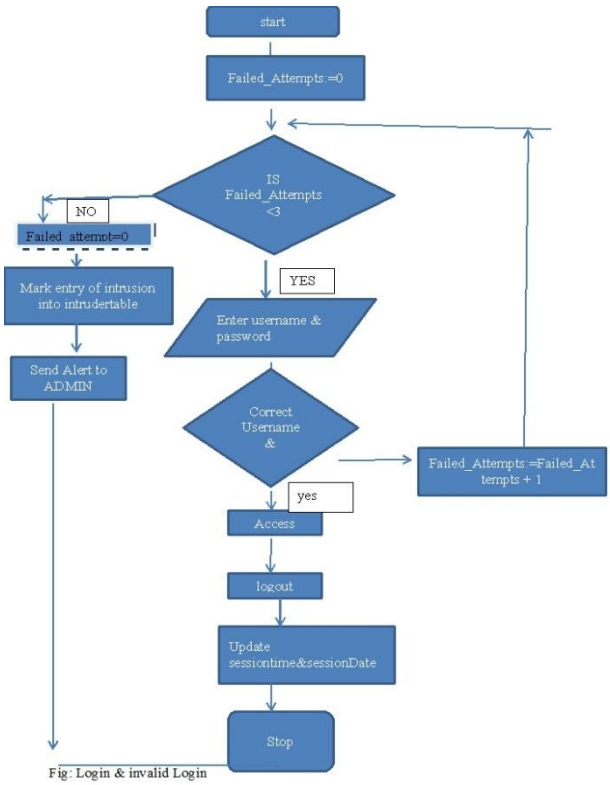

Fig: Login & invalid Login

yes

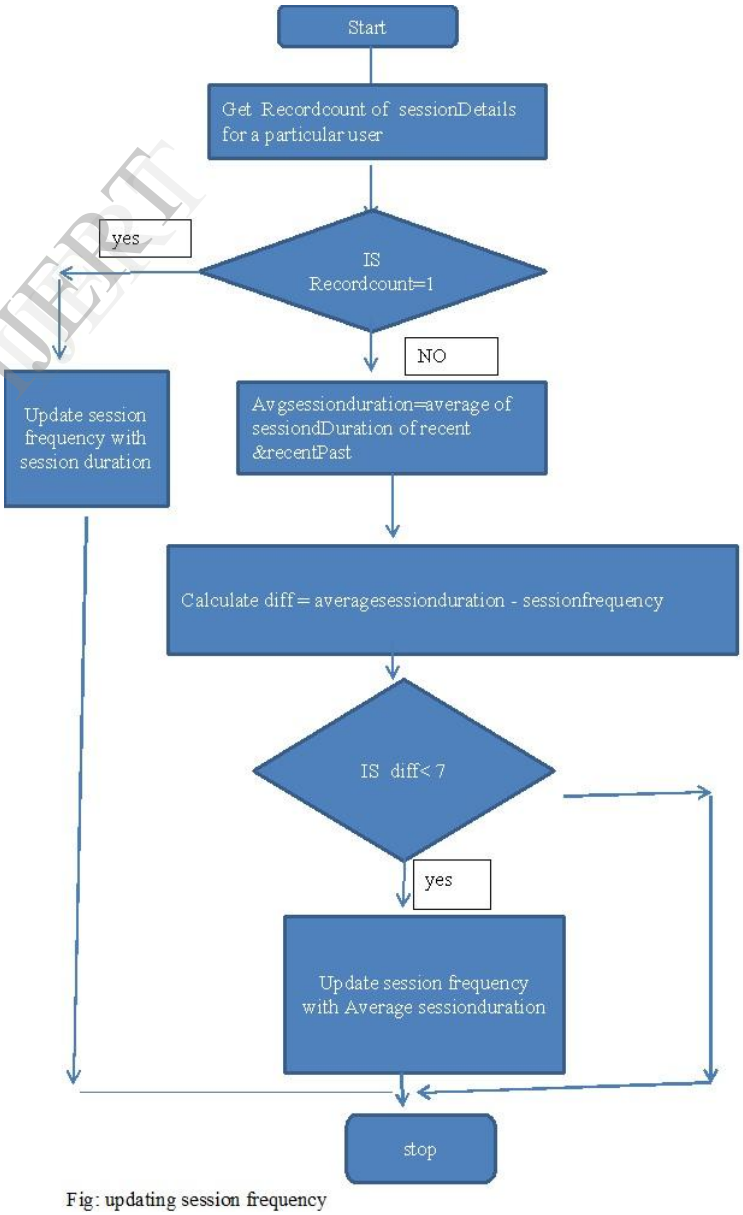

Fig: updating session frequency

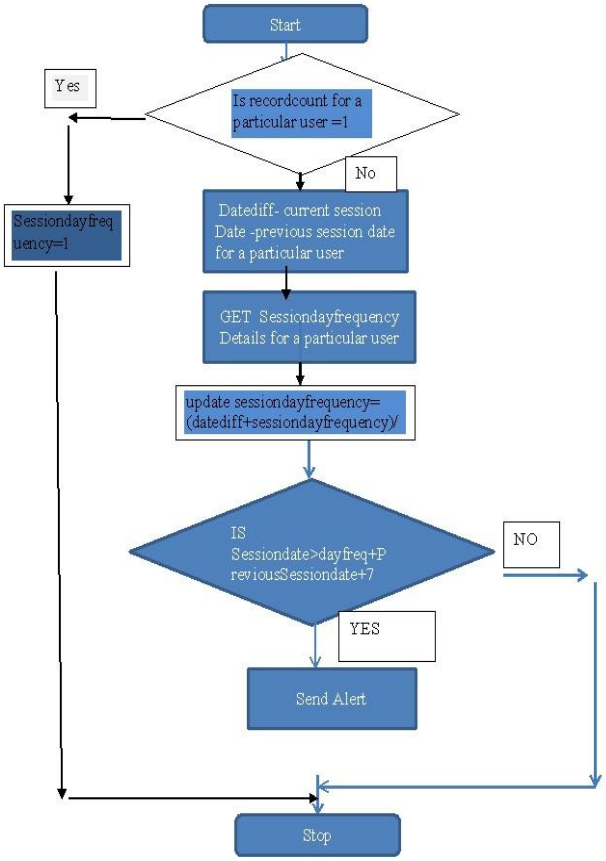

Fig: Check for Session day frequency

IV. EVALUATION

Sample of hundreduser's data was tested for the designed application. An alert in the form of email was sent to Database Administrator in the case of invalid login trail and unusual usage of application by the users.

A simple banking application was designed to perform basic operations. The record of session day frequency and session frequency and its Updating scenario is shown below.

Session Date	Session Duration (seconds)
4thjune 2013	10
14thjune 2013	20
16th june2013	5
30thaug 2013	30

	Day Frequency	Session Frequency
Initial scenario	1	10
Latest Scenario	39	17

SCENARIO	ALERT
Three trails of login failed	email
Correct Username and Password but different frequency of typing from routine	email
Frequency of session either in terms of number of days or duration is Unusual	email

V. CONCLUSIONS

Even though this technique of concept-drift and feature-evolution is efficient for novel class detection but still there are some drawbacks:

1. False alarm rate, i.e. detecting existing classes as novel is high for some data set.
2. If there is more than one novel class we are unable to distinguish among them.

So we propose flexible decision boundary for outlier detection by allowing a slack space outside the decision boundary. This space is controlled by a threshold and the threshold is adapted continuously to reduce the risk of false alarm and missed novel classes. By studying the pattern of intrusion we can spot new protocols for design of application

VI. REFERENCE

[1] Sanjay Chawla. Feature selection, association rules network and theory building. In The 4th Workshop on Feature Selection in Data Mining, 2010.
[2] M. Dash and H. Liu. Feature selection for clustering. In Proceedings of 4th Pacific Asia Conference on Knowledge Discovery and Data Mining, 2000.Springer-Verlag, 2000.
[3] Jennifer G. Dy and Carla E. Brodley. Feature selection for unsupervised learning. J. Mach. Learn. Res., 5:845–889, 2004. ISSN 1533-7928.
[4] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufman, 2001.
[5] H. Liu and H. Motoda. Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic Publishers, 1998. ISBN 0-7923-8198-X.
[6] H. Liu and H. Motoda, editors. Computational Methods of Feature Selection.Chapman and Hall/CRC Press, 2007.
[7] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. on Knowledge and Data Engineering, 17(3):1–12, 2005.
[8] H. V. Nguyen and V. Gopalkrishnan. Feature extraction for outlier detection in high- dimensional spaces. In The 4th Workshop on Feature Selection in Data Mining, 2010.
[9] Zenglin Xu, RongJin, Jieping Ye, Michael R. Lyu, and Irwin King. Discriminative semi- supervised feature selection via manifold regularization. In IJCAI' 09: Proceedings of the 21th International Joint Conference on Artificial Intelligence, 2009.
[10] B. Babcock, S. Babu, M. Datar, R. Motawani, and J. Widom. Models and issues in data stream systems. In PODS, 2002.
[11] S. Guha, N. Milshra, R. Motwani, and L. O'Callaghan. Clustering data streams. In FOCS, pages 359–366, 2000.