A FRAUD DETECTION USING ONE-TO-MANY DATA LINKAGE OF ONE CLASS CLUSTERING TREE

Vikram G. Lachake[1], Prakash P. Rokade[2]
[1] *M.E.(Computer Engg.) II, Computer Engineering, SND COE & RC, yeola, Maharashtra, India*
[2] *Head Of Department, Information Technology, SND COE & RC, yeola, Maharashtra, India*

ABSTRACT

There is need to compare records in one data set with records in another data set due to increased awareness in many countries of the potential of record linkage for fraud detection. Record linkage is traditionally performed among tables to cluster the data. The proposed method aims to perform One-to-many data linkage for fraud detection i.e. to associate one record in Table A with one or more matching records in Table B based on OCCT tree. The OCCT tree provides One-to-many record linkage between objects of same or different types. It is easy to build OCCT tree and convert into linkage rules. The inner nodes of OCCT tree contains attribute from table A and the leafs holds a compact representation of a subset of records from Table B which are more likely to be linked with record from Table A, whose values are according to the path from the root of the tree to the leaf.. The OCCT tree is induced small amount of using splitting and pruning methods and contains nodes to avoid over fitting. Old methods are taken long time for one-to-many linkage. The OCCT based on One Class approach that is it considers only positive examples (matching examples). Hence the proposed method provides better performance in terms of precision and recall as compared to C4.5 decision tree-based linkage method.

Keyword: - *Fraud Detection, Dataset, One Class Clustering Tree, Data Deduplication, Record Linkage.*

1. INTRODUCTION

Fraud detection is to identify masquerade attack or identity deception. Record linkage of given two data sets used to form training data to identify fraudulent users using probabilistic model. Record Linkage is the problem of recognizing records in two data sets that refers to same object or entries. The main objective of record linkage is to combine the data sets that do not share common attribute or foreign key. Data set is a collection of records. Record linkage usually performed to reduce to reduce large data set into smaller data set. Data deduplication to remove repeated records is important task in data linkage process. This improves the complexity of system. The data deduplicton is important task in data cleaning processes because duplicates can affect the outcome of subsequent data processing or data mining processes. The record linkage generally performed among identical entities. Record Linkage classified into: One-To-One and One-to-Many record linkage. In One-to-One record linkage, where each object in one data set is potentially linked to single matching object in another data set. In One-to-Many record Linkage, where each object in one data set is potentially linked to group of matching objects in another data set. In this paper, the proposed method performs One-to-Many record linkage using One Class Clustering Tree (OCCT). A clustering tree is a tree in which each of the leaves contains cluster whereas normal tree consist of single

classification. Each cluster is characterized by set of association rule. The proposed method characterized by two things. First it performs One-to-many data linkage between objects of same or different types. Second we use positive examples (matching examples) i.e. One Class approach. This is one of the advantages because obtaining negative examples is difficult task. In fraud detection, it is easy to obtain positive examples. The OCCT can be used in three different domains like fraud detection, data leakage prevention and recommender system. The fraud detection detects the transactions which are not performed by genuine users but by fake users. The recommender system presents users with items in which they are interested. The recommender system works as classification problem in which classifier determines whether users likely to like a certain item. The data leakage prevention detect the anomalous access to database that do not conform to normal access or database misuse. The goal is to link a set of records, representing the perspective of the request, with a set of records representing the data that can be genuinely retrieved within the specific perspective. [1]

2. LITERATURE SURVEY

I.P. Fellegi and A.B. Sunter presented mathematical model within a context of which linkage rules developed to determine records in two different data sets link or non-link to provide guidance for handling of linkage problem. The linkage rules assigns the probabilities for taking each of the three actions i.e. link, non-link or possible error. They defined two types of error as error of decision. First unmatched linked records are actually unmatched and second non-linked records are actually matched [2].

A.J.Strokey, C.K.I. Williams, E. Taylor and R.G. Mann presented One-to-Many record linkage based on Expectation Maximization algorithm. They use the Expectation Maximization algorithm to compute the probability of a given record pair being match and to learn the characteristic of matched records. The method is derived for specific astronomical problem of far-infrared observations to optical counterpart, but is generally applicable. They described theory of record linkage but does not discuss its application or its implementation [6].

M.Yakout, A.K.Elmagarmid used entity behavior to decide two different entities are in fact same. Entity's behavior extracted from transaction records. The goal is to merge the behavior of two possible matched entities and determine the gain in behavior pattern as their matching score [3].

M.D.Larsen and D.B.Rubin used maximum likelihood learning amongst candidate models. The maximum likelihood defines some similarity measure between records in one data set and those in another data set. We can use maximum likelihood to classify potential record pairs as either match or non-match [5].

P.Christen and K.Goiser examined parameter-free technique for data linkage for comparison with those techniques of data linkage with parameters. For this they used three string comparison methods (JW, ED and Compress Comparison method) and three classification methods (Decision tree, K-means, Farthest First). They compared three decision tree which are built using three different string comparison techniques. Their work of data linkage limited to one or two attributes and attributes are predefined. Hence this method is difficult to generalize [7].

F.De Comite, F.Denis, R. Gilleron and F.Letouzey introduced POSC4.5 algorithm for record linkage of positive and unlabeled examples. They considered binary classification and hence this method is not generalized. They require not only the data set but also the information of positive examples out of whole data set. The attraction of their work is that they presented modified entropy formula which that considers weight of positive examples in a given data set. They assumed that negative examples are in unlabeled data set as per given distribution [4].

H.Blockeel, L.D. Raedt and J.Ramon presented a top down induction of decision tree in which each leaves contains cluster instead of single classification. Each cluster is characterized logical expression representing records belonging to it [8].

G. Alan Wang, Hsinchun Chen, Jennifer J. Xu, and Homa Atabakhsh presented a technique to automatically detect identity deception using adaptive detection algorithm that suits to incomplete identities with missing values and to large data sets containing millions of records. The authors describe three experiments to show that the algorithm is significantly more efficient than the existing record comparison algorithm with little loss in accuracy. It can identify deception having incomplete identities with high precision. In addition, it demonstrates excellent efficiency and

scalability for large databases [12].

Malek Ben Salem and Salvatore J. Stolfo proposed technique for masquerader attack to identify theft. They used user search behavior that to detect deviations indicating a masquerader attack. They showed that modeling user search behavior reliably detects all masquerader attacks with a very false positive rate than prior work [13].

3. IMPLEMENTATION DETAILS

3.1 Work Breakdown Structure

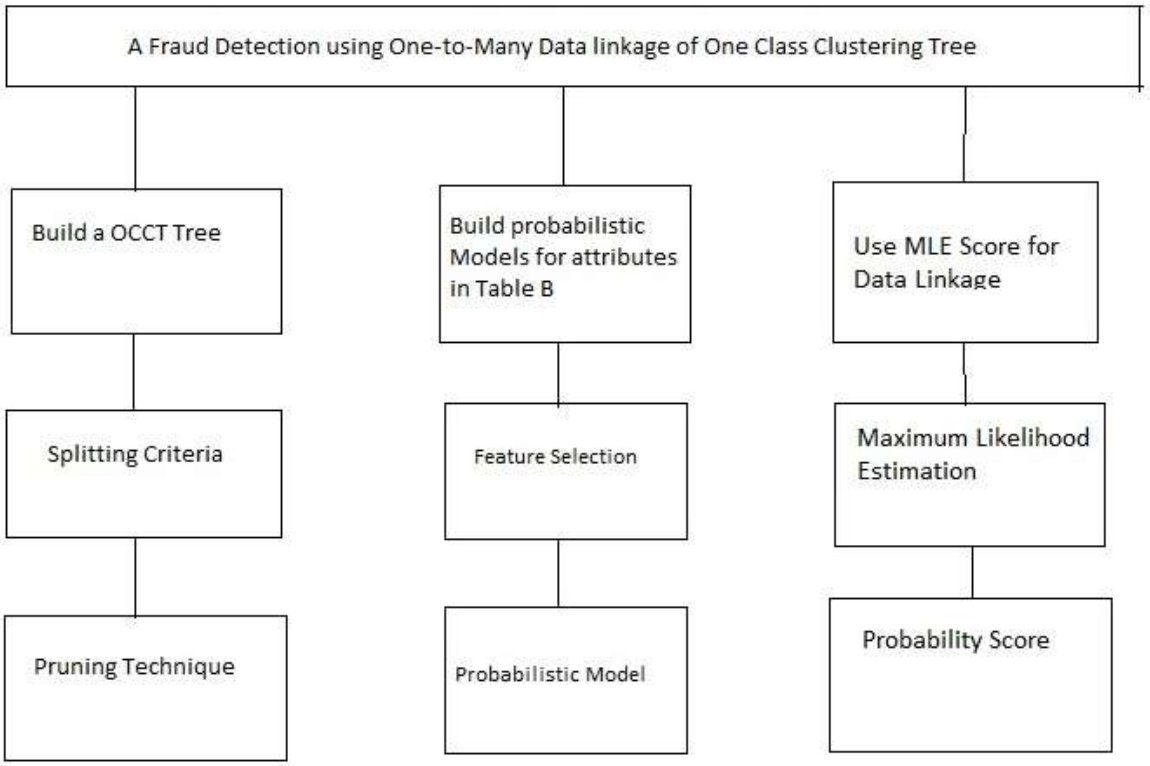

Fig -1: Work Breakdown Structure

The work breakdown structure mainly divided on following areas

- **Build a OCCT Tree**
- **Build probabilistic Models for attributes in Table B**
- **Use MLE Score for data linkage**

3.1.1. Build a OCCT Tree

The One Class Clustering Tree (OCCT) builds using matching examples only. The clustering tree inducing model captures the knowledge of which records are expected to match. The induction of linkage model includes developing the structure of the tree. To build the clustering tree requires selecting the attribute at each level of the tree. The internal nodes of the tree consists of attributes from Table A only. The attributes are intermediate level of tree selected by using one of the splitting criteria. The splitting the attribute used at each step of building the clustering the tree. The splitting criteria ranks the attributes based on how they good are in clustering the matching examples.

In addition a pre-pruning process is implemented. This means that tree stops expanding a branch whenever branch does not improve the accuracy of the model. The pruning process decide which branches to be trimmed.

3.1.1.1 Splitting Criteria

During inducing the clustering tree is that it should contain smallest number of nodes. By reducing the size of the tree (number of nodes) that performs well on training set. It is believed that small tree would better generalize, avoids the over fitting and forms simpler representation for human eye which is easy for human eye to understand. The proposed method will use four splitting criteria to evaluate the splitting of the tree based on attribute of Table A. Each splitting criteria is used to calculate the similarity between two record sets T1 and T2 and is indicated by sim(T1,T2). The splitting criteria that is used determine the attribute that creates the best split of a table that is Table T divided into two Tables TA and TB, which differ from each other as much as possible. Each attribute in Table TA is evaluated to determine the record that it achieves.

3.1.1.2 Maximum Likelihood Estimation (MLE)

The maximum likelihood i.e. probability score is calculated for the attribute which is not selected as splitting attribute by giving the value of other attribute. The attribute having highest maximum likelihood score is selected as the next splitting attribute [10]. The complexity of this method depends on the size of input data set, maximum likelihood score calculated for number of attribute and method used to build or model the tree (e.g. decision tree)

3.1.1.3 Pruning

Pruning is the process to trim unnecessary branches to improve accuracy of model. Thus tree is induced using matching examples only. The pruning process is use to give compact representation of tree i.e. it contains small number of attributes. It also avoids over fitting and improve the time complexity. There are two types of pruning process 1) Pre-pruning and 2) Post-pruning.

The pre-pruning work on top down approach i.e. the pruning process is done during the tree induction process when further split does not give complete knowledge of record matching. linkage model.

The post-pruning work on bottom up approach i.e. the pruning process done after completion of inducing linkage tree when further split does not give complete knowledge of record matching. In our proposed system we are using pre-pruning process to reduce the time complexity. The decision whether to prune the branch taken once best splitting attribute is chosen. We propose either MLE or LPI our system. In Maximum Likelihood Estimation (MLE), a MLE score is calculated for each of splitting attribute. If none of the candidate attribute achieve MLE score greater than current node then branch is pruned and current node becomes leaf node.

3.1.2 Build probabilistic Models for attributes in Table B

Once the tree is built then each leaf contains the matching record from Table B. The probabilistic model is built for each attribute of Table B by giving the values of other attributes. There are two goals for this step. First is to reduce the size of tree by produce the compact representation of tree and to avoid overfitting. It is not necessary to create probabilistic model for each attribute of Table A. The attributes having specific meaning in leaf, the models are created for those attributes only. These attributes are selected using feature selection process. The purpose of feature selection process is to best represent of records in leaf. Let there are | B | number of attributes indicating the records from given table B. B = b1, b2, b3,.....b | B |. For each attribute bi in table TB, a probabilistic model Mi is built by giving the values of other attributes b1, b2, b3 etc.

3.1.3 Use MLE Score for data linkage

In this linkage step, Maximum Likelihood Estimation (MLE) is calculated foe each possible pair of records. The MLE score indicate the probability of record pairs are being match. The cardinality of record pairs are multiply by MLE score. Then MLE score is compared with given threshold to decide given record pairs are match. If MLE score of record pair is greater than threshold then record pairs are categorized as match otherwise it is categorized as non-match.

Algorithm: Fraud Detection using One-to-Many Data Linkage of One Class Clustering Tree

Input: Set of records from Dataset 1and Dataset 2, Set of attributes from Dataset 1 and Dataset 2 and threshold value.

Output: Set of matching records from Dataset 1 and Dataset 2 to identify fraudulent users.

Method Begins

Step 1: Load Dataset 1 and Dataset 2.

Step 2: Apply preprocessing on Dataset 1 and Dataset 2.

Step 3: Create training set from two input Datasets.

Step 4: Calculate the probability for each attribute.

Step 5: Based on probability calculate the MLE score.

Step 6: Based on best MLE score select root of tree.

Step 7: After that select next level attribute of tree based on based on threshold value to construct tree.

Step 8: Check if MLE based pre-pruning technique applied at each level of tree. If not? Apply.

Step 9: Apply decision model based on MLE score to match records from to two Datasets to identify fraudulent users.

Method Ends

4. MATHEMATICAL MODEL

Suppose I1 and I2 denotes datasets of which records are to be match , O denotes the matched record set, F1 denotes the function that accomplish OCCT tree building, F2 denotes the function that accomplish representation of leaf using probabilistic model, F3 denotes the function that accomplish linkage of records using OCCT tree. Venn diagram shown in fig.2 denotes the mapping of input to output.

I = {I1,I2}

F = {F1,F2,F3}

O = {O1,O2,....On}.

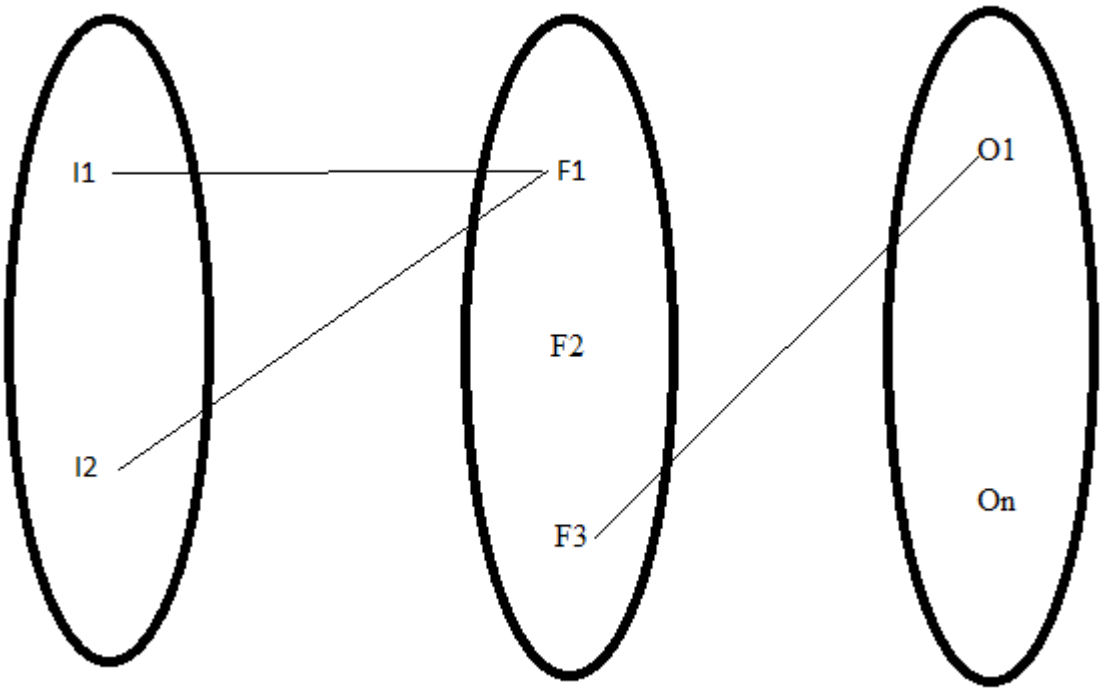

Fig -1: Venn diagram

5. RESULTS AND DISCUSSION

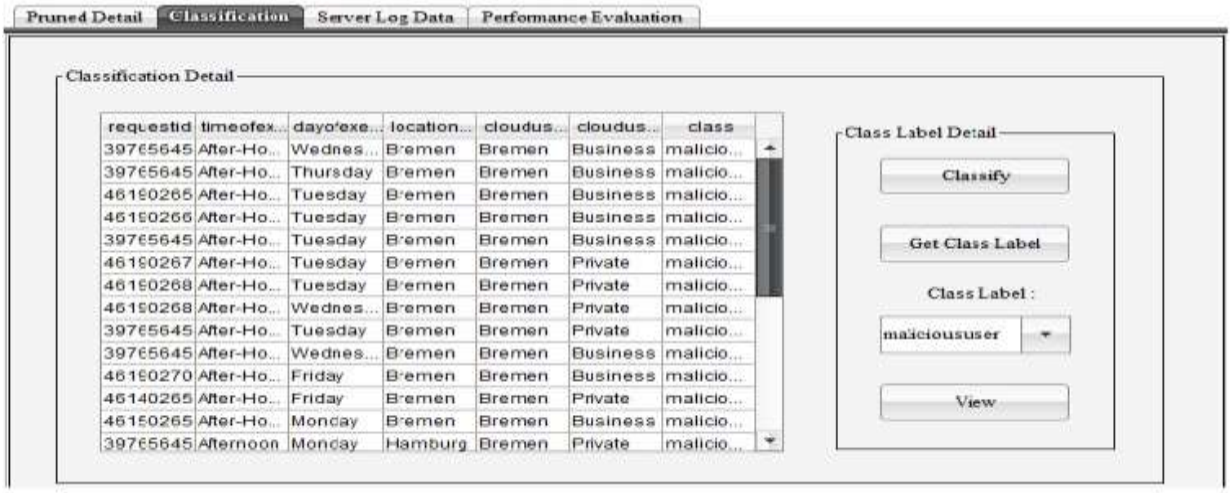

Fig -2: Fraud detection showing malicious users using One-to-Many data linkage of OCCT

As shown in fig-2 the proposed system we implemented fraud detection using One-to-Many data linkage of One Class Clustering Tree. The transactions performed by the entire user are used as training data for our system by taking Cartesian product of two tables context.data and customer.data. Finally we got all transactions containing fraudulent activity.

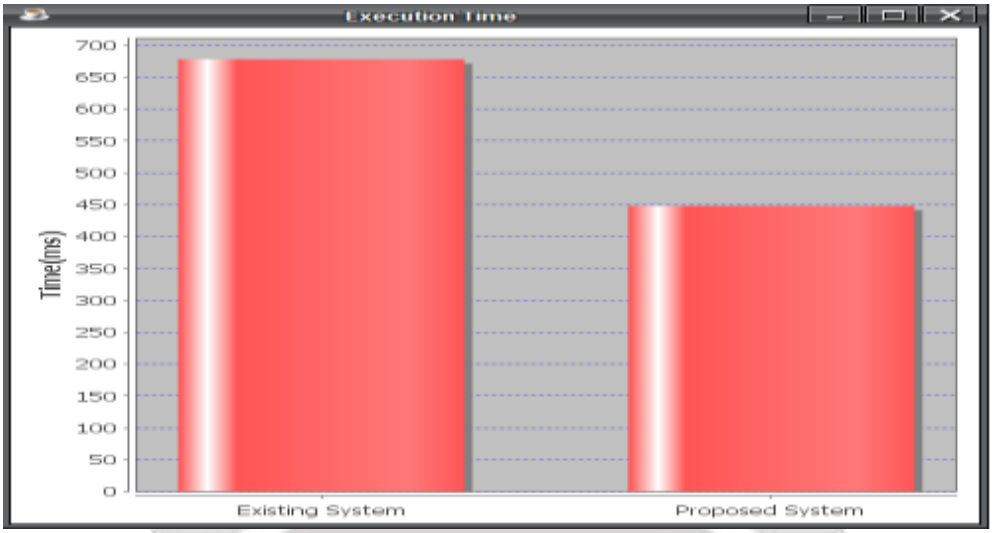

Fig -3: Comparison of proposed system with existing system.

Fig-3 shows the execution time comparison of fraud detection using One-to-Many data linkage of One Class Clustering Tree with existing system. Hence proposed system is efficient in execution time.

6. CONCLUSIONS AND FUTURE SCOPE

In this paper we showed how fraud detection can be performed by using One-to-Many data linkage of dissimilar entities. It is possible to link two data sets that does not share common attribute i.e. foreign key. This system useful for fraud detection including to identify identity deception or masquerader attack using One-to-Many data linkage of One Class Clustering tree . The work can be extended to fraud detection using Many-to-Many data linkage of One Class Clustering Tree.

7. ACKNOWLEDGEMENT

I am very much thankful to my respected project guide Prof.P.P.Rokade, for his motivation and guidance proved to be very helpful important during the development of this dissertation work. I am also thankful to our P.G.Coordinator Prof.V..N.Dhakane for supporting us while preparing for dissertation work. I am also thankful to our Head of Department Prof.Shaikh I.R. I would like to thank all the faculties who have helped me during my dissertation work. Lastly I am thankful to all my friends who shared their knowledge in this field with me.

8. REFERENCES

[1]. Maayan Dror, Asaf Shabtai, Lior Rokach, and Yuval Elovici, OCCT: A One-Class Clustering Tree for implementing One-to-Many Data Linkage, IEEE Transactions on Knowledge and data Engineering, Vol. 26, No. 3, March 2014.

[2]. I.P. Fellegi and A.B. Sunter, A Theory for Record Linkage, J. Am. Statistical Soc., vol. 64, no. 328, pp. 1183-1210, Dec. 1969.

[3]. M. Yakout, A.K. Elmagarmid, H. Elmeleegy, M. Quzzani, and A. Qi, Behavior Based Record Linkage, Proc. VLDB Endowment, vol. 3, nos. 1/2, pp. 439-448, 2010.

[4]. F. De Comite, F. Denis, R. Gilleron, and F. Letouzey, Positive and Unlabeled Examples Help Learning, Proc. 10th Intl Conf. Algorithmic Learning Theory, pp. 219-230, 1999.

[5]. M.D. Larsen and D.B. Rubin, Iterative Automated Record Linkage Using Mixture Models, J. Am. Statistical Assoc., vol. 96, no. 453, pp. 32-41, Mar. 2001.

[6]. A.J. Storkey, C.K.I. Williams, E. Taylor, and R.G. Mann, An Expectation Maximization Algorithm for One-to-Many Record Linkage, Univ. of Edinburgh Informatics Research Report, 2005.

[7]. P. Christen and K. Goiser, Quality and Complexity Measures for Data Linkage and De-duplication, Quality Measures in Data Mining, vol. 43, pp. 127-151, 2007.
[8]. H.Blockeel, L.D.Raedt, and J.Ramon, Top down Induction of Clustering Tree, ArXiv computer Science e-prints, pp 55-63,1998.
[9]. D.J. Rohde, M.R. Gallaghar, M.J.Drinkwater, and K.A.Pimpplet, Matching of Catalogues by Probabilistic Pattern Classification, Monthly Notices of the Royal Astronomical Soc., vol. 369, no. 1, pp. 2-14, May 2006.
[10]. D.D.Dorfmann and E.Alf, Maximum-Likelihood Estimation of Parameters of Signal Detection Theory and Determination of Confidence Intervals-Rating-Method Data, Journal of Math Psychology, vol. 6, no. 3, pp. 487-496, 1969.
[11]. S.Guha, R.Rastogi and K.Shim, ROCK: A Robust clustering algorithm for Categorical attributes, Information System, Vol.25, no.5, pp. 345- 366, July 2000.
[12]. G.A. Wang, H. Chen, J.J. Xu, and H. Atabakhsh, "Automatically Detecting Criminal Identity Deception: An Adaptive Detection Algorithm," IEEE Trans. Systems, Man and Cybernetics, Part A: Systems and Humans, vol. 36, no. 5, pp. 988-999, Sept. 2006.
[13]. M.B. Salem and S.J. Stolfo, "Modeling User Search Behavior for Masquerade Detection," Proc. 14th Symp. Recent Advances in Intrusion Detection, 2011.