ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2014

Applying Modified K-Nearest Neighbor to Detect Insider Threat in Collaborative Information Systems

Aruna Singh[1], Smita Shukla Patel[2]

P.G. Student , Department of IT, SKNCOE, Pune University, India[1]

Asst. Professor, Department Of IT, SKNCOE, Pune University, India[2]

ABSTRACT: Collaborative information systems have acquired a lot of attention recently by providing all the information at one place. These systems can be used in all scenarios where there are many user roles defined and a lot of common information is accessed by them. In such cases, a huge possibility of threats from insiders exists. This is due to the fact that users have access to all the subjects irrespective of their roles. Users may sometimes misuse the system by taking out the information for some invalid reasons. It is very difficult to avert such situations. The work proposed here provides a way out of detecting such anomalous activity by making us of patterns of usage and a modified k nearest neighbor algorithm. The proposed work does not require any type of access control mechanism or extra information about the users or subjects. It is purely dependent on the access log of the users which is automatically generated once the user accesses the subjects. The relational patterns of access logs are analyzed for nearest neighbors in terms of number of subjects accessed as well as metainformation related to those subjects. Deviation is calculated for all the users. Anomalous users show larger deviation from their nearest neighbors. The proposed work improves the accuracy of the algorithm by adding few more parameters of validity and weight while calculating the deviation. It is proved by the experiments that the detection of anomalous users is more likely in case of modified nearest neighbor algorithm.

KEYWORDS: Anomaly Detection, CAD, CIS, k nearest neighbor, threat.

I. INTRODUCTION

Multiple users can share common resources in collaborative information systems. In such systems there is a chance of having similar level of access for all users. But the larger scope of access rights also results in encouraging illegitimate as well as immoral use of information. As a result much work has been done to provide proper access control which helps in providing security from outsider threat. Role based access control or experience based access control is also implemented to gain more flexibility. Insider threat detection has evolved much lesser compared to the outsider threat detection techniques. It is also difficult to detect, as the threats are authorized users who are performing some activities which are not genuine. The insider threat detection mainly uses the anomaly detection methods because of causing anomalous behaviour. Anomaly is understood as a pattern in the data that does not behave as expected. There are techniques in data mining too for this type of threat detection. These are of two types supervised learning and unsupervised learning. Supervised learning makes use of training set data which has to be available in advance. It has many problems e.g. necessity of labelled data and inability to detect rare events. In contrast unsupervised learning methods do not require any such prior information. Its success only depends on the proper selection of similarity measures, feature selection etc. This approach is very helpful in detecting rare events called as outliers or exceptions by calculating the deviation from expected behaviour.Generally there are two types of solutions which consider the insider threat. The first one uses access rules within the system to control the user's activities. The second method involves the

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2014

review of patterns in which the user has behaved. When there is a large organization, even if there are access rights provided, still the information can be gathered and leaked in a wrong way.

II. RELATED WORK

A recent work done to detect the insider anomaly makes use of a framework called as community anomaly framework detection (CADS) which infers the user communities from the access logs of the users. It then uses the nearest neighbour algorithm to find the deviation from other patterns. Any unusual deviation points to detection of insider threat [10]. Specialised network anomaly detection is a work implemented earlier. It tries to create patterns or communities based on the access of various users on one subject. It then checks whether the community pattern with and without that user is sufficiently different. If there is a remarkable difference than earlier community, it points out that the corresponding user is creating an anomaly. This approach creates a model based on the subject and the users who try to access that particular subject. This model works under the hypothesis that if a user is accessing a particular subject for some anomalous reason than a genuine one, it will definitely mismatch with the existing pattern and its similarity with the network will be much lesser than the other genuine users. This approach is termed as specialized network anomaly detection because it only considers the local view of the system i.e. it considers only one subject and the users around it. The system does give an idea of the complete view of the network as a whole including all the users and all the subjects. The specialized network anomaly detection method gets its required information from the access logs of the users. This system has basically two components:

1) Similarity Management

2) Anomaly Evaluation [12]

Another work dealt by making use of role based access control and experience based access management. In this work, the concept called as “role prediction” is developed [8]. This concept refers to the ability to learn whether a particular user satisfies a particular role due to its access pattern. This concept is very useful for role engineers who need to think about the various roles within an organization. Rather than thinking on their own, it proves comparatively easier to take some related data of usage and analyze it for extracting roles. It also acts as a valuable tool in dedicating assignments to the extracted roles. Also it helps with the situations where two roles are being confused upon. In such cases the roles can be merged. It also relates to experience based access management as roles are being extracted on the basis of experienced data. It is very helpful in dynamic environments where it is not easier to predict the roles at one time only. The work has major contribution as Role Classification, Intelligent role abstraction, Empirical Evaluation. There is another related implementation which makes use of the access transactions. It is implemented to review the already existing access control policies by making use of auditing access logs [2]. It helps in assessing the existing policies as well as figuring out some unknown behavior if present. This implementation also has two steps as follows:

1) Network Construction

2) Association Rule Discovery

All the above mentioned systems used some techniques and representations to make the anomaly detection easier. There have been previous works based on graph based anomaly detection [9]. It becomes very much easier to find out the anomalies if the relationship is depicted as a graph. Each vertex or edge in a graph corresponds to some relationship between the users and subjects accessed. In one of these implementations the idea is that a graph substructure is showing an anomaly if it is not isomorphic to the graph’s normative substructure by X%. In such cases X shows the percentage of vertices and edges which need to be added/removed so as to make the graph isomorphic. There are basically three categories of graph anomalies:

1) Insertions: These constitute the presence of an unexpected edge or vertex which has to be removed.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2014

2) Deletions: These constitute a missing edge or vertex which has to be added to fix the graph.

3) Modifications: These can constitute both of the above as well as changing the vertices of an edge.

Another concept implemented is that of incremental outlier detection. In incremental outlier detection the techniques identifies the anomaly as soon as it takes place. This approach is also used for activity monitoring. An incremental outlier detection algorithm depends on the numbers of nearest neighbors surrounding it. The work further proved that insertion of a new data or deletion of the same affects to only the few neighbors which are very close to it. Thus it makes clear that any insertion or deletion operation does not affect the no of updates required [3].

The graph based approaches generally implemented bipartite graphs. A bipartite graph can be partitioned in a disjoint manner in two sets [5]. Such graphs make it very simple to find out the association rules. For example traders vs. stocks etc. Such graphs can support the operations for similar neighbor identification as well as anomalous insider detection.

Statistical methods observe the activities of subjects and the profiles are generated to represent the behavior. The profile consists of various measures as distribution of access of records as well as cpu utilized during a particular transaction. This work implements by basically keeping two profiles as the previously stored profile as well as the current profile. Such intrusion detection systems regularly update the current profile and after a particular duration compare it with the stored profile. An anomaly score is calculated which is the difference between the current profile and the stored profile. A threshold is already decided for detecting the anomaly. If the calculated anomaly score reaches that score or is higher, then instantly an anomaly is detected [1].

Another insider detection is based on scenarios of anomalous behavior, various temporal sequences and the unfamiliar graph evolution etc. There are various algorithms and representations for dynamic graph processing. This was done in view of scaling well with the business level requirements and deployments in real time data streams [7].

III. OVERVIEW OF ANOMALY DETECTION FRAMEWORK

The proposed system for insider anomaly detection will used the CADs framework with an extra parameter as weight and a modified anomaly detection algorithm which will also consider it [6]. The weight will be calculated depending upon the communities. As the basic k-nearest neighbor algorithm uses only the Euclidian distances between the neighbors for comparing the users, it might not prove sufficient for detecting anomalies .Adding an extra parameter for the nearest neighbor algorithm will make it more stronger and accurate. Fig. 1 gives an overview of this framework.

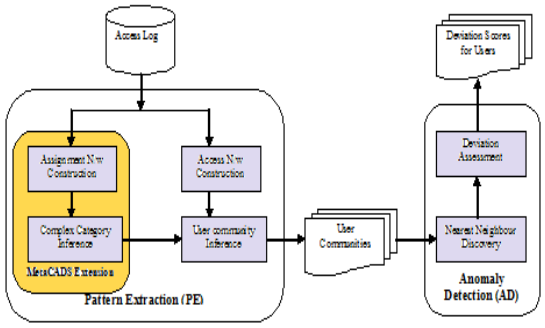

Fig. 1. Anomaly Detection Framework

CADS & METACADS

The community anomaly detection system CADS framework doesn't depend on any prior knowledge of the roles of users or any other extra information for detecting the anomalies [11]. It has two basic steps which not only cover anomaly detection but also the processing of data required prior to anomaly detection. These are

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2014

1) Pattern Extraction

2) Anomaly detection

COMMUNITY EXTRACTION

As the users access various subjects in a collaborative information system, a bipartite graph is created depicting the user subject access. Based on the graph, communities are extracted. The users accessing the similar subjects fall under the same community. The extracted communities are helpful in anomaly detection. The metaCADS framework includes two more steps of assignment network construction and complex category inference. It also considers the semantics of the subjects accessed. Fig. 2. illustrates both of them.

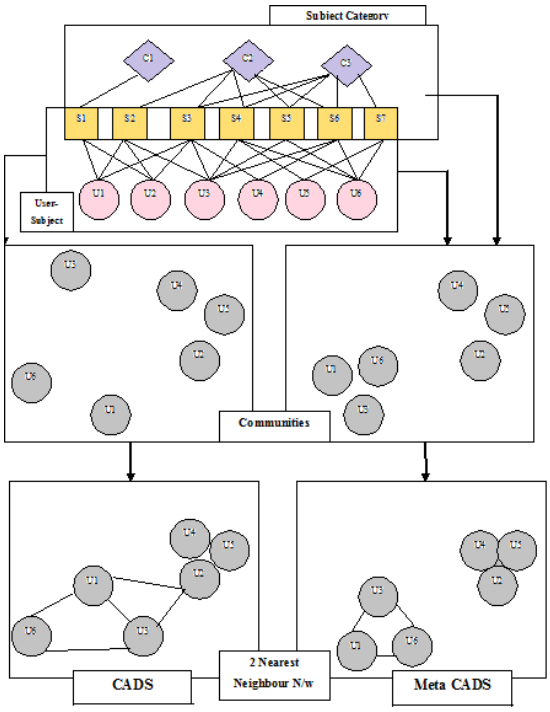

Fig. 2. CADS and MetaCADS

ANOMALY DETECTION

Anomaly detection compares the behaviors of the users to the communities inferred earlier. This is based on the K nearest neighbor algorithm which implements the idea of comparing deviation of each user with its nearest neighbor. If there is any user which does not come under any community, then it is considered to be anomalous.

The proposed system functions are explained as follows:

1) GRAPH BASED PATTERN EXTRACTION:

Graphical Representation and Network Construction: It consists of the depicting the access log in the form of a tripartite graph i.e. user – subject and subject – category graph. The user – subject relationship is the access network and the subject – category relationship is the assignment network. The information in these networks is summarized in matrices.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2014

Community Extraction: The community extraction is done separately for access network and assignment network. Communities are formed based on the similarities of subjects accessed and also similarity of categories of subjects accessed.

2) ANOMALY DETECTION BASED ON MODIFIED K NEAREST NEIGHBOR ALGORITHM

Discovering the user's k nearest neighbor is implemented first. It is based on the algorithm which minimizes the network community profile (NCP)[4].The modified k nearest neighbor algorithm makes use of an additional element i.e. weight which is based on the validity of the user under inspection and the distance deviation within the nearest neighbors. Validity is a parameter which compares the particular user's similarity with its k nearest neighbors based on the class label. For example for a user x

$$validity(x) = \frac{1}{k}\sum_{i=1}^{k} S(lbl(x), lbl(N_i(x)))$$

Here $N_i(x)$ is the *i*th nearest neighbor of x.

Also S is defined as follows:

$$S(a,b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases}$$

In the proposed system the class label will be the community label which is determined when the communities are formed in the above mentioned step of pattern extraction. Validity along with deviation is used in calculating weight as follows:

$$W(i) = Validity(i) \times \frac{1}{d_e + 0.5}$$

Here d_e is the deviation of the nearest neighbour.

This combination of validity with distance based deviation makes the algorithm stronger and minimises any distance based inaccuracies. Threshold values are decided for deviation as well as weights which are considered by the system for deciding the anomalous user.

RESULTS

It is seen from the experiments that both the k nearest neighbour algorithm and modified k nearest neighbour algorithm are showing some unusual deviation for a particular user in the user deviation graph shown in fig. 3. But the deviation is more prominent in case of modified knn due to the added factor of weight and validity. Also ,it is seen from the deviation chart that cad provides consistent results in both the cases whereas Metacad gives unexpected results in some cases. Graph for modified knn for cad very clearly indicates the anomalous user activity.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2014

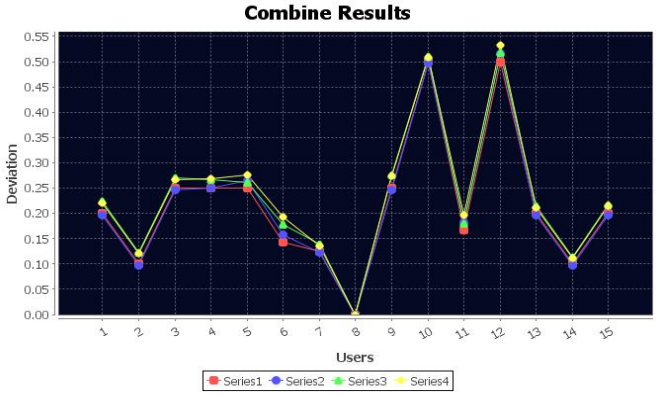

Fig. 3. Distribution of User Deviation

IV. CONCLUSION

In this paper community anomaly detection is taken forward by making use of a modified k nearest neighbor algorithm. The only information required to detect the anomalies is the access log of the users. The modified algorithm has shown more promising results as it has successfully made the deviations more distinct. It is clear from the results that the modified algorithm proves to be stronger than the algorithm used earlier and at times it detects anomalies which might not get detected in previous work. Still a lot of work can be done in future. Deciding the threshold value is a task which requires a lot of data and deviations for consideration. There can be cases of false positives as exceptions. Also manual intervention of experts might be required to confirm the anomalies and the steps to be taken further to stop it. The above mentioned system works for masqueraders but if an anomalous user is acquainted to the system he might work according to some genuine pattern and my go unnoticed. All this can be taken as a future enhancement for this work.

REFERENCES

[1] A. Patcha and J. M. Park, 2007, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," Computer Networks, 51(12):3448–3470.
[2] B. Malin, S. Nyemba and J. Paulett, 2011, "Learning Relational Policies from Electronic Health Record Access Logs," J. Biomedical Informatics, vol. 44, no. 2, pp. 333-342.
[3] D. Pokrajac, A. Lazarevic and L. Latecki, 2007, "Incremental Local Outlier Detection for Data Streams," Proc. IEEE Symp. Computational Intelligence and Data Mining, pp. 504-515.
[4] J. Leskovec, K. Lang, A. Dasgupta and M. Mahoney, 2008 "Statistical Properties of Community Structure in Large Social and Information Networks," Proc. 17th Int'l Conf. World Wide Web, pp. 695-704.
[5] J. Sun, H. Qu, D. Chakrabarti and C. Faloutsos, 2005, "Neighborhood Formation and Anomaly Detection in Bipartite Graph," Proc. IEEE Fifth Int'l Conf. Data Mining, pp. 418-425.
[6] P. Hamid, A. Hosein and M. B. Behrouz, "MKNN: modified k-Nearest Neighbor," Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, 2008, pp. 831-834.
[7] T. E. Senator, H. G. Goldberg, A. Memory, W. T. Young, B. Rees, R. Pierce, D. Huang, M. Reardon, D. A. Bader, E. Chow, I. Essa, J. Jones, V. Bettadapura, D. H. Chau, O. Green, O. Kaya, A. Zakrzewska, E. Briscoe, R. I. L. Mappus, R. McColl, L. Weiss, T. G. Dietterich, A. Fern, W. Wong, S. Das, A. Emmott, J. Irvine, J. Lee, D. Koutra, C. Faloutsos, D. Corkill, L. Friedland, A. Gentzel and D. Jensen, 2013, "Detecting insider threats in a real corporate database of computer usage activity," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, pp. 1393-1401.
[8] W. Zhang, C. Gunter, D. Liebovitz, J. Tian and B. Malin, 2011, "Role prediction Using Electronic Medical Record System Audits," Proc. Ann. Symp. Am. Medical Informatics Assoc., pp. 858-867.
[9] W. Eberle and L. Holder, 2009, "Applying Graph-Based Anomaly Detection Approaches to the Discovery of Insider Threats," Proc. IEEE Int'l Conf. Intelligence and Security Informatics, pp. 206-208.
[10] Y. Chen and B. Malin, 2011, "Detection of anomalous insiders in collaborative environments via relational analysis of access logs," Proceedings of the first ACM conference on Data and application security and privacy, pp. 63-74.
[11] Y. Chen, S. Nyemba and B. Malin, May-June 2012, "Detecting Anomalous Insiders in Collaborative Information System," Dependable and Secure Computing, IEEE Transactions on, vol.9, no.3, pp. 332-344.
[12] Y. Chen, S. Nyemba, W. Zhang and B. Malin, 2011, "Leveraging Social Networks to Detect Anomalous Insider Actions in Collaborative Environments," Proc. IEEE Ninth Intelligence and Security Informatics, pp. 119-124.