A Novel Approach in Clustering Supervised Learning (CSL) for Recommender Systems

K. A. Balasubramaniam

Bharathiar Univiersity, Coimbatore, and
Department of Computer Science
Arulmigu Kalasalingam College of Arts & Science
Krishnankoil, Tamilnadu, India – 626 126.
Email: kabala1977@gmail.com

M. Chidambaram

Department of Computer Science
Rajah Serfoji Government College
Thanjavur, Tamilnadu, India – 613 005.
Email: chidsuba@gmail.com

***Abstract* – A recommender system is a novel approach in clustering supervised learning for identifying the necessity to develop a recommender system for movies. Here we try to adjust the training and testing samples to get the best accuracy for the recommender system. A recommender system applies the method of knowledge discovery for a specific problem to adapt the recommendation of the products and services to the users. The vast growth in the information and the count of visitors to web sites especially on e-commerce applications in the past few years has created some challenges for recommender systems. E-commerce recommender systems are vulnerable to profile injection attacks, involving insertion of fake profiles into the system to influence the recommendations to users. Prior work shows the performance of systems is affected by a small number of biased profiles. In this paper, we propose that a supervised clustering approach can be used effectively for the detection of profile injection attacks in recommender systems. We formulate the problem as a mapping model between rating behavior and item distribution by exploiting the least-squares approximate solution. Experimental results are given to validate the superior performance of our approach in comparison with benchmarked methods and user methods.**

Keywords - clustering, recommender system, performance measure, supervised approach.

I. INTRODUCTION AND BACKGROUND

Numerous web sites attempt to help users by incorporating a recommender system that provides users with a list of items and/or web pages that are likely to interest them. Content-based filtering and collaborative filtering are usually applied to predict the recommendations. Among these two, Collaborative filtering is the most common approach for designing e-commerce recommender systems [1]. It works by building a database of items with users' opinion on them. Then a specific user is matched against the database in order to find the neighbors, where he or she shares similar tastes. As the system is open to user input, chance of attack is more. Researchers have discussed different types of attacks [2]. The ultimate target of all type of profile injection attacks is either to push a product or to nuke a product (or a group of products). In case of Random Attack a pre-specified rating is assigned to the target items and random ratings are assigned to the filler items whereas in average attacks, rating of each filler item corresponds to the mean rating of that item [3].

Some additional attack types have been specified namely Bandwagon Attack Segment attack, Reverse Bandwagon Attack and Love/Hate Attack. The last one is a very simple attack and requires no system knowledge where the attack profile consists of minimum/ maximum rating value for target items and maximum/ minimum rating value for filler items for nuke/push attack [2]. On the other hand, in literature, researchers have proposed several outlier detection techniques [4]. They can be broadly categorized into different groups namely distance based approach, density based approach, clustering based approach and depth based approach. In clustering based approach, the clusters having small number of members are considered as the clusters consisting of outliers assuming that outliers are a small percentage of the total data. At the same time, the nature of the attack data also differs from the data without attack. Based on these two assumptions, the data members of clusters with small sizes are considered as outliers, which in turn, correspond to the attack data [5]. As mentioned, the attack profiles are highly correlated and at the same time the number of attack profiles is very small compared to total number of genuine user profiles. Keeping these two points in view, we have considered the problem of profile-injection attack detection as a problem of outlier detection in the user rating dataset and applied the *Partition Around Medoids,* PAM, clustering algorithm in detecting the outliers or in injecting attack profiles [6], [8].

A. Partition Around Medoids, PAM

The PAM algorithm is intended to find a sequence of objects named medoids that are centrally located in clusters [7]. Objects that are tentatively defined as medoids are placed into a set S of selected objects. If O is the set of objects then the set U = O − S is the set of unselected

objects. The goal of the algorithm is to minimize the average dissimilarity of objects to their closest selected object. Equivalently, we can minimize the sum of the dissimilarities between object and their closest selected object.

In our paper, after evaluating the performance of PAM clustering algorithm in detecting attack-profiles with different size of filler items, an incremental version of the PAM algorithm has been applied to test whether a new user profile which is going to be inserted into the system is an attack profile or not. Then we have applied a PAM based outlier detection algorithm to find attack profiles in large clusters that are not identified by the PAM algorithm. Finally, an angle-based outlier detection strategy is used for finding attack profiles in the attacked database [9].

B. Recommender Systems Used in CSL

A recommender system is an automated software system which can be trained to make decisions intelligently for new and future inputs. The training is given to the intelligent software system using prior events and their outputs which exists in the world history as facts now. Once the system is trained, the users can give inputs for the future problems and the system will do some data mining operations to give the best suited results [10]. In this research paper, we compare the results of Movielens database system obtained after applying data mining techniques like Clustering and Classification and try to create a recommender system to rate movies(high and low) by the users in the scale of 1 to 5 where 1 means poor and 5 means excellent. Once this recommender system is fully established then it can intelligently recommend the consumers whether a movie is viewable or not [11].

II. RELATED WORK

Detection of profile-injection attacks on recommender systems have been studied by many researchers. Supervised classification techniques have been used in order to distinguish attack profiles from genuine user profiles. Researchers have compared the performance of several hierarchical clustering algorithms [12, 13]. They have proposed a two-stage approach of outlier detection by using the concept of minimum spanning tree along with clustering. A Fuzzy-based clustering algorithm has been proposed for detecting outliers in data. PAM has been used as clustering algorithm. A separation technique is used after applying PAM algorithm and small clusters are generated by the clustering algorithm are identified as outliers [14].

The unsupervised methods have been used by many researchers in the field of network intrusion detection. Also in the area of recommender system, a few works have been reported in the literature where unsupervised methods are used as a tool for attack detection [15]. Principal Component Analysis (PCA) based clustering algorithm is used for detecting attack profiles based on the assumption that attack profiles are very highly correlated with each other.

III. CLASSIFICATION AND CLUSTERING METHODLOGY

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of predicting the class of object where class label is unknown. The derived model is based on the analysis of set of training data (i.e., data objects whose class label is known) [16]. The derived model may be represented in various forms such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. Hence the output of the classification depends upon the types of classifier used. In our case, we have used probabilistic approach to find the classification of likes and dislikes of movies, which is represented in terms of probabilities of success or failure [17].

There exist various clustering methods due to the method of finding similarity between clusters as some use cosine methods or Euclidean distance measures or Manhattan distance measures resulting in to clusters of variety of shapes [18]. Basically clustering can be done in two broad-spectrum traditions:

A. Partitioning Approach of Clustering

Here we assume that all tube clustered item in one big cluster initially. Then we start verdict similarity between the data-items and forms sub-clusters and further repeat the process of verdict similarity with in each of those sub-clusters until we get a cluster with most similar item-sets [19, 20]. Here we use Top-Down approach for clustering called as divisive approach of clustering in hierarchy.

B. Agglomerative Approach of Clustering

Here we assume that all item sets are individual to find similarity between item-sets and merge those item sets to form a root [21, 22]. Again these roots are used to find similarity amongst them and other un-clustered data-items (also called clusters) and again superior roots are formed. This approach is called Agglomeration approach [23, 24].

IV. EXPERIMENTS AND RESULTS

In this section, the performance of our proposed method is evaluated with existing techniques. To validate the effectiveness of deep learning based detection system a MovieLens-100k is used in comparison with benchmark method KNN which is illustrated in Fig 1. To be notified that, the detection performance of ours outperforms than KNN for bandwagon Attack with different filler size of (2%,3%,6%,8%,10%). Moreover, Fig 2 illustrates the

segment attack detection for the different filler size improves the performance of KNN method and no change occurs when attacker size is increased.

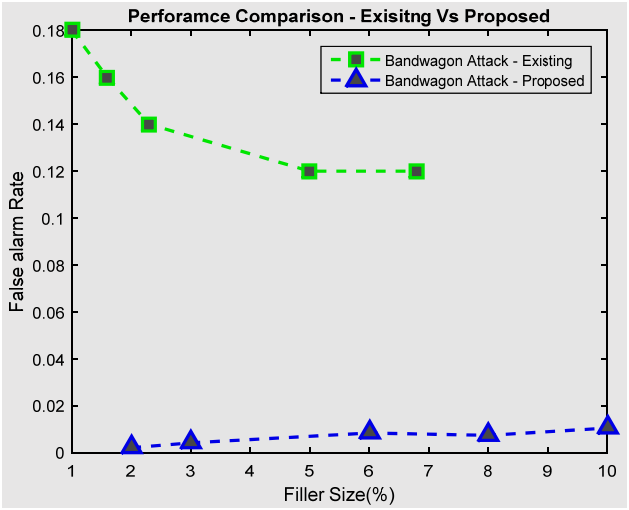

Fig 1. The comparison of false alarm rate with different filler size for Bandwagon attack detection

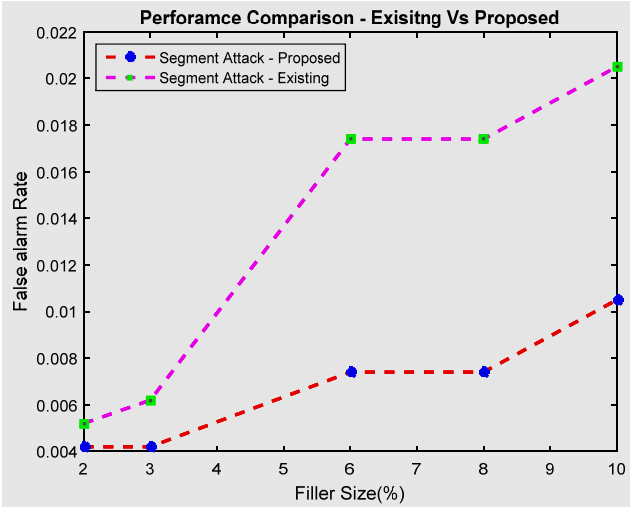

Fig 2. The comparison of false alarm rate with different filler size for Segment attack detection

The next observation Fig3 and Fig 4 shows that the detection rate eventually attains the peak value with the filler size increasing for both bandwagon and Segment attack respectively. But, the detection rate shows minimum value when attack size is small. The results specify that it is difficult to fully solve the imbalanced classification in our method. In addition, the detection rate gradually gets improved and attained a steady rate with increase in the filler size. These results indicate that the proposed features also improve our method to detect Bandwagon and Shilling attacks.

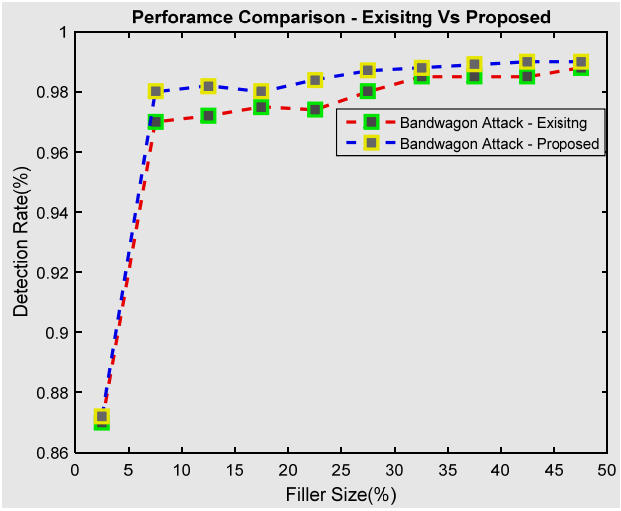

Fig 3. The comparison of Detection rate with different filler size for Bandwagon attack detection

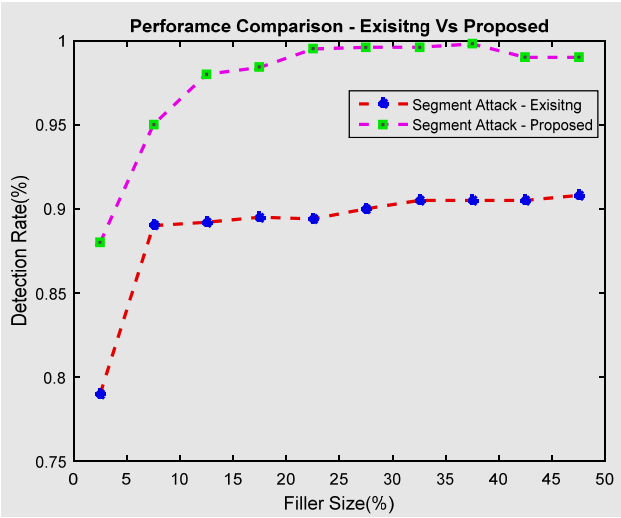

Fig 4. The comparison of Detection rate with different filler size for Segment attack detection

V. SUMMARY AND CONCLUSION

On summarizing the study we can conclude that test results for movie lens Recommender System gives good result if nearly 80K data is taken for training the system. Also we can use Clustering Approach for identifying the view ability of a movie based on Movie Ratings. Hence, such a Recommender systems, if built with intelligence using machine learning, may bring the future decisions more and more closer to human thinking and may even sometimes go beyond human thinking limits. We provide a narrative detection approach for detecting attacks, which constructs a mapping model by exploiting the relationship between rating behavior and item distribution. Widespread experiments on both the MovieLens-80K and Movie Lens-ml-latest-small datasets demonstrate the effectiveness of the proposed approach. To compare with the benchmark

method, our proposed method shows more optimistic detection performance. In future, we propose to extend and improve attack detection in the following directions: Upward a theoretical grounded work method for the detection problem is also of interest and an extracting more simpler and effective features from user profiles.

REFERENCES

[1] Lam, S. And Riedl, J. Shilling recommender systems for fun and profit. In Proceedings of the 13th International WWW Conference (New York, NY)(2004).

[2] Burke, R,Mobasher, B.,Williams, C., And Bhaumik, R. 2006b. Detecting profile injection attacks in collaborative recommender systems. In Proceedings of the IEEE Joint Conference on Ecommerce Technology and Enterprise Computing, E-Commerce and E-Services (CEC/EEE 2006, Palo Alto, CA)(2006).

[3] Mehta Bhaskar, 2007. Unsupervised Shilling Detection for Collaborative Filtering. Association for the Advancement of Artificial Intelligence (www.aai.org), 2007.

[4] Loureiro,A., L. Torgo and C. Soares, 2004. Outlier Detection using Clustering Methods: a Data Cleaning Application, in Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany.

[5] John Peter.S., Department of computer science and research center St.Xavier˝s College, Palayamkottai, An Efficient Algorithm for Local Outlier Detection Using Minimum Spanning Tree, International Journal of Research and Reviews in Computer Science (IJRRCS), March 2011.

[6] Cutsem, B and I. Gath, 1993. Detection of Outliers and Robust Estimation using Fuzzy Clustering, Computational Parthasarathi Chakraborty and Sunil Karforma / Procedia Technology 10 (2013) 963 – 969 969 Statistics & Data Analyses 15, pp. 47-61.

[7] Acuna E. and Rodriguez C., (2004), A Meta Analysis Study of Outlier Detection Methods in Classification, Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, available at academic.uprm.edu/~eacuna/paperout.pdf.

[8] Al- Zoubi, M. B., An Effective Clustering-Based Approach for Outlier Detection, European Journal of Scientific Research, Vol. 28, No. 2, 2009, pp. 310-316.

[9] Eskin,E., Arnold, A., Prerau,M., Portnoy, L., Stolfo, S.: A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data. In: Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publichsers Inc, pp. 255-262. (2000).

[10] Portnoy, L., Eskin, E., Stolfo,S.: Intrusion detection with unlabeled data using clustering. In: Proceeding ACM Workshop on Data Mining Applied to Security. (2001).

[11] Mehta Bhaskar, 2007. Unsupervised Shilling Detection for Collaborative Filtering. Association for the Advancement of Artificial Intelligence (www.aai.org), 2007.

[12] M. O. K. Bryan and P. Cunningham, "Unsupervised retrieval of attack profiles in collaborative recommender systems," in Technical Report, University College Dublin, 2008.

[13] MacQueen, J.,1967. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. and Prob, pp. 281-97.

[14] Kaufman, L. and P. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.

[15] Chakraborty, P.S., "A Scalable Collaborative Filtering based Recommender System using Incremental Clustering", In: Proceeding of IEEE International Advance Computing Conference, 2009.

[16] H.-P. Kriegel, M. S. hubert, and A. Zimek. Angle-based outlier detection in high dimensional data. In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 444–452, New York, NY, USA, 2008. ACM.

[17] Mishra R., Modi N., "A Novel Approach to cluster new data-items in previously clustered data-items using Agglomerative Clustering with Single Link.

[18] Jiawei Han and Micheline Kamber. Data Mining: concept and Techniques. Elsevier Publication.

[19] Jai Prakash Verma, Sapan Mankad, "Smart Inbox: A comparison based approach to classify the incoming mails", International Journal of Artificial Intelligence and Knowledge Discovery Vol.1, Issue 1, Jan, 2011.

[20] Indian E-Commerce Stats: Online Shoppers & Avg Order Values To Double In Next 2 Years, Web Link" http://trak.in/tags/business/ 2014/04/04/indian-e-commerce-growth-stats/f"on dated 20-06-2015.

[21] Weka 3: Data Mining Software in Java", Web Link: http://www.cs.waikato.ac.nz/ml/weka/ on dated 20-06-2015.

[22] W. N. Venables, D. M. Smith and the R Core Team, "An Introduction to R", Notes on R: A Programming Environment for Data Analysis and Graphics Version 3.2.0 (2015-04-16).

[23] Jai Prakash Verma, Bankim Patel, Atul Patel, "Big Data Analysis: Recommendation System with Hadoop Framework", 2015 IEEE International Conference on Computational Intelligence & Communication Technology, 978-1-4799-6023-1/15, 2015 IEEE DOI 10.1109/CICT.2015.86.

[24] Sandra Garcia Esparza, Michael P. O'Mahony, Barry S myth, "Contents lists available at Sci Verse Science Direct Knowledge-Based Systems", Knowledge-Based Systems 29 (2012) 3–11.