Literature review

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Paper | objective | Method | Results | Advantage | Disadvantage | Different from our study |
| [1] | the safeguarding of electronic patient record (EPR) systems in particular | presents a system that employs  a density-based local outlier detection model. Patterns in EPR data are extracted to profile user behavior and device interactions  in order to detect and visualize anomalous activities.  It considers the relative density of  points and can detect data in biased datasets  LOF  employs the relative-density of a coef\_cient against its neighbours  as the indicator of the degree of the object being an  outlier  The LOF anomaly score measures the  local deviation of density through determining how isolated  the value given by *k-*nearest neighbours (k is set to 5).  A value below  1 indicates a dense region, and would therefore also be an  inlier. A value signi\_cantly above 1 therefore indicates an  outlier (anomaly). As all values within the range 0-1 are  classi\_ed as inliers, values within the range 1-2 were also  classi\_ed as inliers. Any value above 2 was considered to  indicate an outlier for the purposes of this experiment | The system is able to detect 144 anomalous behaviors  in an unlabeled dataset of 1,007,727 audit logs. | It considers the relative density of  points and can detect data in biased datasets. This means  that it is advantageous over proximity-based clustering. LOF  employs the relative-density of a coef\_cient against its neighbours  as the indicator of the degree of the object being an  outlier | The resulting values are [quotient](https://en.wikipedia.org/wiki/Quotient)-values and hard to interpret. A value of 1 or even less indicates a clear inlier, but there is no clear rule for when a point is an outlier. In one data set, a value of 1.1 may already be an outlier, in another dataset and parameterization (with strong local fluctuations) a value of 2 could still be an inlier. These differences can also occur within a dataset due to the locality of the method.[ https://en.wikipedia.org/wiki/Local\_outlier\_factor] | Other studies were not accessed in this method |
| [2] | this paper focuses on detecting privacy breaches resulting from inappropriate  accesses of EHRs. | collaborative filtering inspired approach to predicting inappropriate accesses  We analyze the access violation problem in terms of *dyadic prediction* [8], where the goal is  to predict a label for the interaction of a pair of entities. This framework captures scenarios  such as recommending friends in a social network [21], predicting student performance on  test scores [19], and clickthrough rate prediction in computational advertising [14].  *Collaborative filtering* is a popular strategy for this problem,  where one attempts to tease out the implicit characteristics of users and items based solely  on these historical preferences, and use these to predict preferences for every (user, item)  pair.  This is similar to an item recommendation problem: we are trying to measure the affinity a  particular (user, patient) pair has for violation.  he underlying assumption of the collaborative filtering approach is that if a person *A* has the same opinion as a person *B* on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person | **The CF out performed SVM, liniear and logistic regression on AUROC, AUPRC and G-Mean respectively** |  | In practice, many commercial recommender systems are based on large datasets. As a result, the user-item matrix used for collaborative filtering could be extremely large and sparse, which brings about the challenges in the performances of the recommendation.  One typical problem caused by the data sparsity is the [cold start](https://en.wikipedia.org/wiki/Cold_start_(recommender_systems)) problem. As collaborative filtering methods recommend items based on users' past preferences, new users will need to rate sufficient number of items to enable the system to capture their preferences accurately and thus provides reliable recommendations.  Similarly, new items also have the same problem. When new items are added to the system, they need to be rated by a substantial number of users before they could be recommended to users who have similar tastes to the ones who rated them. The new item problem does not affect [content-based recommendations](https://en.wikipedia.org/wiki/Content-based_filtering), because the recommendation of an item is based on its discrete set of descriptive qualities rather than its ratings.  **Scalability**[[edit](https://en.wikipedia.org/w/index.php?title=Collaborative_filtering&action=edit&section=13)]  As the numbers of users and items grow, traditional CF algorithms will suffer serious scalability problems[[*citation needed*](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)]. For example, with tens of millions of customers {\displaystyle O(M)} and millions of items {\displaystyle O(N)}, a CF algorithm with the complexity of {\displaystyle n} is already too large. As well, many systems need to react immediately to online requirements and make recommendations for all users regardless of their purchases and ratings history, which demands a higher scalability of a CF system. Large web companies such as Twitter use clusters of machines to scale recommendations for their millions of users, with most computations happening in very large memory machines.[[15]](https://en.wikipedia.org/wiki/Collaborative_filtering#cite_note-twitterwtf-15)  **Synonyms**[[edit](https://en.wikipedia.org/w/index.php?title=Collaborative_filtering&action=edit&section=14)]  [Synonyms](https://en.wikipedia.org/wiki/Synonyms) refers to the tendency of a number of the same or very similar items to have different names or entries. Most recommender systems are unable to discover this latent association and thus treat these products differently.  For example, the seemingly different items "children's movie" and "children's film" are actually referring to the same item. Indeed, the degree of variability in descriptive term usage is greater than commonly suspected.[[*citation needed*](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)] The prevalence of synonyms decreases the recommendation performance of CF systems. Topic Modeling (like the [Latent Dirichlet Allocation](https://en.wikipedia.org/wiki/Latent_Dirichlet_Allocation) technique) could solve this by grouping different words belonging to the same topic.[[*citation needed*](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)]  **Gray sheep**[[edit](https://en.wikipedia.org/w/index.php?title=Collaborative_filtering&action=edit&section=15)]  Gray sheep refers to the users whose opinions do not consistently agree or disagree with any group of people and thus do not benefit from collaborative filtering. [Black sheep](https://en.wikipedia.org/wiki/Black_sheep) are a group whose idiosyncratic tastes make recommendations nearly impossible. Although this is a failure of the recommender system, non-electronic recommenders also have great problems in these cases, so having black sheep is an acceptable failure | Only 4 methods (CF,SVM, Linear and logistic) were compared |
| [3] | detect atypical behavior in a health care application. | The J48 algorithm in WEKA produces a C4.5  decision tree. |  | First this model scales to large datasets – in the  examples detailed below; the system is processing tens of thousands of instances. Secondly a decisions tree has the benefit of  producing output that can be easily interpreted by a user and it could be easily understood why a given instance was  misclassified [4]. | The first method  uses ad-hoc analysis to find atypical behavior by visual inspection looking for interesting nodes. Ie either ad-hoc or path-length investigation | There was no comparison for a better choice of algorithm |
| [4] | introduce a community-  based anomaly detection system (CADS), an unsupervised  learning framework to detect insider threats based on infor-  mation recorded in the access logs of collaborative environ-  ments. Comparison of different anomaly detec-  tion methods on the EHR dataset. The number of  accessed subjects for simulated user is random. CADS is based on the observation that typical users  tend to form community structures, such that users with  low affinity to such communities are indicative of anoma-  lous and potentially illicit behavior. The model consists of  two primary components: relational pattern extraction and anomaly detection | k-nearest neighbors (KNN), PCA and CADS: In essence, CADS is a hybrid of KNN and  PCA. In the CADS-AD component, the behaviors of the users  in the CIS access logs are compared to the community pat-  terns. Users that are found to deviate significantly from  expected behavior, as prescribed by the patterns, are pre-  dicted as anomalous users. As in the CADS-PE component,  the CADS-AD component consists of a process to translate  access log transactions into scored events. | Figures 11, it can be seen that the performance  of the volume model in this setting is poor.  in Figures 11 and 12. It can be  seen that CADS exhibits the best performance of simulated  user detection (according to AUC). At the lowest mix rate,  CADS was almost two times more accurate at the most spe-  cific tuning level. Moreover, CADS is only marginally af-  fected by the mix rate, whereas the other approaches are  much more sensitive. |  |  | Only fur methods were compared but in our study, various classification methods were compared which include …. |
| [5] | The goal is to use statistics, machine learning,  knowledge of workflows and other techniques to support an  anomaly detection framework that finds such users  In this paper  we introduce and study a random topic access model or RTA aimed  at users whose access may be illegitimate but is not fully random  because it is focused on common semantic themes | Latent Dirichlet  Allocation (LDA), for feature extraction, a k-nearest neighbor (k-  NN) algorithm for outlier detection and evaluate the ability to  identify different adversarial types. |  |  |  | The reason for the choice of combination of these methods were unknown. |
| [6] | to detect anomalous user  behaviors based on the sequence of their requests within a web  session. We first decompose web sessions into workflows based on  their data objects. In doing so, the detection of anomalous sessions  is reduced to detection of anomalous workflows | hidden Markov model, Distance-based model were compared | THMM performed better than the distance model |  |  |  |

In general, anomalies can be defined as any observations that are different from the normal behavior of the data[7]

1. Boddy, A.J., et al., *Density-Based Outlier Detection for Safeguarding Electronic Patient Record Systems.* IEEE Access, 2019. **7**: p. 40285-40294.

2. Menon, A.K., et al., *Detecting inappropriate access to electronic health records using collaborative filtering.* Machine learning, 2014. **95**(1): p. 87-101.

3. Ziemniak, T. *Use of Machine Learning Classification Techniques to Detect Atypical Behavior in Medical Applications*. in *2011 Sixth International Conference on IT Security Incident Management and IT Forensics*. 2011.

4. Chen, Y., et al., *Specializing network analysis to detect anomalous insider actions.* Security Informatics, 2012. **1**(1): p. 1-24.

5. Gupta, S., et al. *Modeling and detecting anomalous topic access*. in *2013 IEEE International Conference on Intelligence and Security Informatics*. 2013.

6. Li, X., Y. Xue, and B. Malin. *Detecting anomalous user behaviors in workflow-driven web applications*. in *2012 IEEE 31st Symposium on Reliable Distributed Systems*. 2012. IEEE.

7. Das, K., J. Schneider, and D.B. Neill. *Anomaly pattern detection in categorical datasets*. in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008.