Anomaly detection in medical record access logs.

September, 2017

Contact

David Strömberg
david.stromberg@omegapoint.se

1 Introduction

This degree project concerns misconduct detection in medical records access logs with the use of unsupervised machine learning methods. In a preliminary study, an approach has been suggested for the specific data set in question, and the aim of the project is to implement this method, adjust it according to new findings and finally evaluate the result.

2 Background

Medical records are personal and possibly sensitive documents. The Patient Data Protection act[1] decrees that only the users of the medical record system (medical professionals etc.) whom need the records to carry out their profession are legally allowed to access them. However, in the system currently used in Stockholm county[2], the user access in not restricted based on the user-patient relationship. Instead, the access is restricted based on the care-provider[3] system and supplemented by restricted access markers. A user can easily override the restrictions of this system, which means that a user has access to far more patient records than required. To detect and discourage illicit access, each query to the system is registered and information about the access event is stored. This results in huge quantities of data, on the order of 10^8 stored events in total, which need to be analysed to detect possible access fraud.

The fraud detection system in place today consists of a set of parametrised rules which each care-unit[4] can tune for their liking. Each month, these rules are applied to roughly 10% of the users' log data, chosen at random. Any violations are marked automatically, and then audited manually. It is likely that both the precision and coverage of this system can be improved with complementary machine learning techniques. Precision means the number of false-positives, i.e. the instances/users which have violated the specific rule-set without misusing their access rights. Coverage concerns misconduct currently not being detected due to only 10% of the data being analysed each month, or due to false-negatives, i.e. misconduct not covered by the existing rules.

The problems of increasing the precision and that of increasing the coverage have been separated into different projects. This way, different aims can be formulated and different approaches

[1] Author's translation. Swedish: Patientdatalagen
[2] Swedish: Stockholms län
[3] Swedish: Vårdgivare
[4] Swedish: Vårdavdelning

can be taken. In this project, the aim is to improve the coverage, that is, to discover cases of misconduct which are false-negatives in the current rule-based system. Consequently, the main data of interest is unlabelled and unsupervised learning needs to be used. Following a literature survey, a method aimed at achieving increased coverage has been proposed [1]. In summary, the proposed method consists in first clustering the data using either constrained clustering [2] or spectral clustering [3], and then to apply anomaly detection techniques within each cluster. More details can be found in the report [1].

3 Aim and research question

The aim of this degree project is to implement and evaluate the anomaly detection method which have been proposed in [1] for detection of illicit access in patient's medical records.

The aim can also be formulated as a set research questions: How can the method proposed in [1] be implemented as an algorithm for misconduct detection in medical record access log data? Is there an interpretation of the result when the method is applied to the log data? If yes, what is the interpretation? Which advantages or disadvantages does the proposed method hold in comparison to the current rule-based system?

4 Project evaluation

The implementation aim will be evaluated by presenting functioning and well-written code. Functioning means that all the different parts specified in the proposed procedure, and any necessary modifications, are included in the code and that the code has been debugged and tested. Well-written means that the code is structured and easy to read. This way, Omegapoint will be able to reuse or refer to this code in the future.

The evaluation of the output of the algorithm presents a problem since the ground truth, whether the unlabelled data is actually cases of misconduct or not, will not be known. Therefore, the evaluation of the output of the algorithm will be focused on interpretability. If this method is to be used within the health care sector, the users must be able to interpret the results so that a judgement can be made if further investigation is warranted or not.

However, since interpretability is a subjective evaluation criteria, it should also be compensated by using some of the available labelled data as a check. By keeping some of the labelled data out of the training, it can be used to see if the algorithm detects it as anomalous or not. This way, a comparison with the rule-based model can be performed.

5 Proposed procedure

The project can be divided into four phases: pilot study, clustering, anomaly detection and comparison with the rule-based system. Both the clustering phase and the anomaly detection phase will consist of finding an appropriate feature representation of the data, observing the results of the method, and then go back and reformulate the features if needed.

5.1 Pilot study

The literature survey containing the proposed method [1] should be the first material to be read. Attached to [1], and also attached to this specification, are lists of literature on which the proposition has been based. The topics include contextual anomaly detection, constrained clustering, spectral clustering, kNN for outlier detection and more.

5.2 Clustering

It is expected that the clustering phase will be the most time-consuming. First, the access events must be aggregated into suitable month-user representation. Together with the representation, a similarity metric which the clustering algorithms can use has to be chosen. Since the raw data consists of events and contains many categorical features, it is advisable that the month-user representation describes frequencies over the categories (more details in the preliminary study [1]).

Second, either a constrained clustering [2] or a spectral clustering [3] algorithm is applied to the processed data. Each cluster will represent a user-month profile and will be used to put the data into different contexts. If time allows, more than one clustering algorithm could be tested and the results compared. The clustering should be performed several times to determine robustness. If the result is unsatisfactory, a change of the aggregated feature, the similarity metric or both might be required.

It could prove difficult to evaluate the result of the clustering. Ideally, each cluster will have a clear interpretation as a profile (i.e. a user type), but since interpretability is a subjective criteria it will be difficult to evaluate. A useful discussion on objective evaluation criteria can be found on this page[5] [4].

There are constrain clustering and/or spectral clustering algorithms available online (see this repository[6] and in scikit-learn[7]). If an appropriate algorithm cannot be found online it might be necessary to implement one from scratch. In this case, the aim of the project should be reformulated to allow for the required extra time.

5.3 Anomaly detection

Once the profiles/clusters have been found, an anomaly detection algorithm is applied within each cluster and an anomaly score is calculated for each event. Furthermore, the labelled data available could be used to determine if some clusters are more likely to contain cases of misconduct and the anomaly score could be weighted accordingly.

The proposed anomaly detection algorithm is kNN for outlier detection [5]. Details on the implementation are provided in [5], so it should be possible to reimplement in whichever programming language is chosen for the project. Similar to the clustering phase, it will be necessary to find a feature representation and a similarity metric for the access events, and again, this representation may have to be adjusted after observing the outcome of the anomaly scoring.

5.4 Comparison with rule-based system

Once the algorithm is implemented and tested, a comparison can be made with the current rule-based system. As mentioned, could be difficult to find an objective performance metric, so instead, the focus should be on interpretation of the observed differences. For example, which events are marked as suspicious for the respective methods? Is there overlap, or is there a significant difference? Which features are important for the respective methods?

[5] https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html
[6] https://github.com/Behrouz-Babaki/COP-Kmeans
[7] http://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

References

[1] Ciwan Ceylan. Detection of illicit access of medical records using machine learning. Technical report, Omegapoint, 2017.

[2] Sugato Basu, Ian Davidson, and Kiri Wagstaff. *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press, 2008.

[3] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416, 2007.

[4] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[5] Stephen D Bay and Mark Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 29–38. ACM, 2003.