Detection of Illicit Access of Medical Records using Machine Learning

Ciwan Ceylan[1]

Abstract

This is a preliminary study of possible application of machine learning and data mining methods to discover illicit use of Stockholm county's medical record system. The users of the system have access to far more patient records than they are legally permitted to open, and therefore, information about each access event is registered for later analysis. Currently, the analysis is done via a rule-based system which is manually tuned, and supplemented by manual auditing. The aim of this study is to survey the literature and to suggest how the current analysis system could be complemented with the help of machine learning and data mining methods. The ambition is to increase both the precision and the coverage of the misconduct detection.

[1] *Omegapoint, Analyst*

***Contact**: ciwan.ceylan@outlook.com

Contents

1. Motivation and Aims

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 bases on the care-giver[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 aim of this investigation is to lay the groundwork for a future project, i.e. identifying the key difficulties of the task, surveying the available literature and make recommendations for an approach based on the findings.

2. Data Description

For each access event, there is a range of information available, e.g. the patient's ID, the user's ID, the care-unit which was accessed, the age of the patient, whether the user and the patient have a personal relation and whether special access was requested by the user, i.e. if the user requested access to a record which he/she would normally not have access to, and the reason for this special access, e.g. emergency or with the patient's approval. The information assoiciated with a data instance will be referred to as attributes. Many of the mentioned attributes are contextual, i.e. the attribute provides a context in which an anomaly can be detected [1]. An example is the ID of the care-unit which have been accessed. The ID alone does not indicate any unusual activity, but when coupled with the information that the user works at a different care-unit than the one access, suspicions should be raised.

In addition to the log file attributes, there are classification labels available as provided by the current rule-based system. This system classifies each event as either red, orange or green, and while these labels are less interesting in and of themselves, there also exists manual labels on a subset of the

[1] Author's translation. Swedish: Patientdatalagen

[2] Swedish: Stockholms län

[3] Swedish: Vårdgivare

[4] Swedish: Vårdavdelning

red (and possibly orange) events. These manual labels are set by a professional who has determined whether these events were true- or false-positives. Consequently, the manually labelled events are expected to lie on the edge of where the rule system detects true-positives and false-positives. However, not much can be said about possible false-negatives, i.e. the events marked as green by the rule-based system and which therefore have not been manually labelled.

An important design question is how the data should be preprocessed. One alternative is that each access event constitutes a data instance. This way, no information would be lost in during preprocessing, but training a model could be difficult since many of the attributes are categorical, e.g. care-unit ID and special access markers, making individual events difficult to compare and possibly noisy. A natural preprocessing step would be to convert the events to frequencies measured for a time window of one month. These frequencies would represent summaries of user behaviour and enable comparison between users rather than individual access events, likely reducing the noise in the data. Furthermore, by considering each users and month as a data instance, the number of samples would effectively be reduced by 2-3 orders of magnitude, reducing computational complexity. On the other hand, if the wrong summary frequencies are chosen for the user attributes, important information might be lost. For example, an access event marked as ”with the patient’s approval” for a patient below the age of 15 could be difficult to capture in a frequency representation. Some ideas for frequencies which could be calculated are: the total number of some event type, e.g. the total number of emergency access events, the number of different event types, e.g. the number of different care-units accessed, or full distributions, e.g. the distribution of access times across the hours of the day.

Another important decision is whether all data should be used together or be separated based on prior knowledge. In the current system, different rules are applied for different care-units since ”normal” behaviour different from unit to unit. Likewise, it might be a good idea to train a separate model for each care-unit, or even for each user. However, the differences might not be as significant as one would expect a priori. Therefore, an important aim of the data analysis is to determine the importance of manually separating the data.

3. Increasing Precision

The current rule-based system results in several false-positives which must be classified manually. The manual labelling is ultimately needed since professionals have access to more information than is available in the log files, and the suspects should be able to defend themselves. Nevertheless, by applying supervised learning to the manually labelled data, it is possible that the number of false-positives could be reduced. The reason is that machine learning algorithms does not consider one attribute at a time the way a rule-based system does, but also takes linear (and non-linear) transformations of the attributes into account. A flexible discriminatory model might therefore be able to find structure not currently being exploited.

There exists a huge variety of models for supervised learning. Depending on the number of labelled samples, different models should be considered. If the number of labelled samples are large, say more than 10^5, neural networks [2, Chapter 5] could be a good choice. These models are very flexible and therefore require large sample sizes to train. If fewer labelled samples are available, e.g. less than 10^4, then less flexible models such as SVM [2, Chapter 7] could be used. The basic implementation of SVM has quadratic time complexity with the number of samples, but also requires fewer samples to achieve good results compared to neural networks.

Regardless of which model and algorithm is chosen, it is probably more effective to train the model on only the manually labelled data as there might be several false-negatives in the automatically labelled data which could interfere with the training. Moreover, the separation up to the manual labelling can already by achieved with the rule-based system, so there is little point in training a model to do what is already being done.

4. Increasing Coverage

Increased coverage can be achieved by a) increasing the amount of data being processed, and b) not relying on the existing set of rules for classifying the behaviour. It is assumed that none of the false-negatives have been manually labelled. Consequently, a supervised approach does not seem viable, and instead, unsupervised or semi-supervised learning (SSL) approaches are needed. The problem can be thought of as an anomaly detection problem, which relies on the assumption that the false-negatives behave like outliers given the right attributes and contexts [1]. An alternative assumption is that the structure of the false-negatives correlate with the distribution of the true-positives. Under this assumption, a fully semi-supervised approach could be used [3]. It is difficult to determine if these assumptions hold or not, especially since they both rely on a good set of attributes being chosen. In the next section, an approach which aims to use ideas both from anomaly detection and SSL literature is suggested.

Suggested Approach

As mentioned Section 2, several of attributes are contextual. This means that many algorithms cannot be used ”out-of-the-box” as the underlying assumptions may not hold without first treating the data. For example, most anomaly detection algorithms are designed for finding point anomalies [1, 4], and SSL assumes that there is informative structure in the unlabelled data [3]. In [1] there is a general discussion of how to deal with contextual variables in an anomaly detection setting. The approach which seems most suitable for the medical record data is the *profile* approach. The idea is to build behavioural profiles based on either the user’s own history, or by grouping similar users together. These approaches have been use for intrusion detection systems [5], cell-phone fraud detection [6] and credit fraud detection [7].

For each specific application, the details for a fraud detection system differ. Therefore, the details for the illicit access detection system cannot fully rely on previous work. However, based on the profiling idea, the following approach is suggested: First, user profiles are created. For this, each data instance is taken as representing a user for a specific month. The attributes for a user-month instance are different frequencies which summarises the user's behaviour. The profiles are created by clustering [8] the users based on both meta-information and a similarity metric.

Examples of useful meta-information constraints are: all instances related to the same user, but for different months, should belong to the same cluster, and/or that all users with the same profession should belong to the same cluster. These constraints can be incorporated using either constrained clustering [9], e.g. *pair-wise constrained k-means* (PCKmeans) [10], or spectral clustering [11]. K-means clustering methods have a large advantage in having linear time complexity with the number of samples, even when using constraints [10]. Conversely, SS spectral clustering usually have cubic time complexity, which can be reduced to quadratic by using a sparse connection matrix, e.g. an AnchorGraph [12]. An advantage of spectral clustering is that allows for visualisation of the data in low dimensional space, i.e. 2D or 3D, which can be helpful for the analysis.

The choice of similarity or distance[5] metric ultimately depends on the attributes, but the Euclidean and overlap distance are commonly used for numerical and categorical attributes respectively. Alternatively, the Mahalanobis distance could be used for numerical values, meaning that different weights are assigned to each attribute. It can be difficult to set these weights by hand, but a possible alternative is to learn them from the data in a semi-supervised manner [14, 15]. For alternative categorical similarity metrics see [16].

Once the data has been clustered, the profiles can be taken as the mean or median user of each cluster, or some other data summary depending on which similarity metric was used. It is important that the clusters are analysed to see which users have ended up together and to investigate the variation with in each cluster. If there is a very large variation, the profiles might not be very useful and a different approach to the clustering might be needed.

With the profiles established, two different approaches can be taken for discovering cases of illicit access: comparing the users' monthly behaviour with the respective profiles and comparing individual access events within the context of a profile. Two types of alerts could be used for comparing user-month instances with the profiles: either if the user deviates above some threshold from its usual profile, and if the user instance is similar to a profile which contains a high ratio of users which have historically been labelled for misconduct. These approaches correspond to unsupervised anomaly detection and a SS approach respectively.

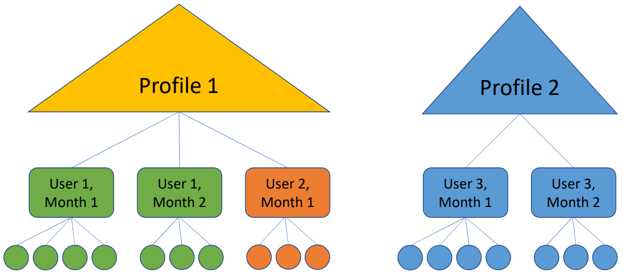

Figure 1. *An illustration of the concept behind the profile system. The user-month instances are clustered to form the profiles. The individual events are only compared within a profile.*

The same distinction between an unsupervised and a SS approaches is found for comparing events within a cluster. Several anomaly detection methods can be found in [1], and a comparative evaluation is carried out in [4]. Of these, *k-nearest neighbours* (kNN) for outlier detection seems appealing. It is a simple method which yet compares well with more advanced alternatives [4]. It does not require training, but testing has quadratic time complexity with the number of samples in a naïve implementation. However, by using a randomisation and pruning trick, the average complexity can be reduced to almost linear, see [17]. kNN further requires a similarity metric, which could be chosen manually, or possibly learned for the data [18]. It is not known if the metric learning can be combined with the pruning technique. For clusters where there are manually labelled events, semi-supervised approaches could be used, many of which are described in [3] and with links to several implementations available at the website detailed by [19].

There are several advantages of the proposed approach. First, it addresses the problem of contextual anomalies by only comparing users and events with and within those profiles which should behave similar, this is the same idea as in *peer group analysis* [20]. Furthermore, it is computationally effective. The initial clustering is only needed once[6], and assigning a new user-month instance to a cluster is cheap since the number of profiles should be relatively few. Furthermore, it enables both comparison of user-month instances and event instances. The system also lends itself to a parallel or distributed implementation since the the comparison of events within the clusters are autonomous. The advantages in scalability is inspired from this paper on anomaly detection in big data from sensors [21].

[5] Note that a distance metric can usually be replaced with a similarity metric. See for example *kernelized k-means* [13].

[6] Or once every year if there is a worry of the data becoming non-representative.

References

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[2] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[3] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. *Semi-supervised learning*. MIT Press, 2006.

[4] Markus Goldstein and Seiichi Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLOS ONE*, 11(4):1–31, 04 2016.

[5] Stephanie Forrest, Steven A Hofmeyr, Anil Somayaji, and Thomas A Longstaff. A sense of self for unix processes. In *Security and Privacy, 1996. Proceedings.*, pages 120–128. IEEE, 1996.

[6] Tom Fawcett and Foster Provost. Activity monitoring: Noticing interesting changes in behavior. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 53–62. ACM, 1999.

[7] Richard J. Bolton, David J. Hand, and David J. H. Unsupervised profiling methods for fraud detection. In *Proc. Credit Scoring and Credit Control VII*, pages 5–7, 2001.

[8] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

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

[10] Sugato Basu, Arindam Banerjee, and Raymond J Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 333–344. SIAM, 2004.

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

[12] Wei Liu, Junfeng He, and Shih-Fu Chang. Large graph construction for scalable semi-supervised learning. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 679–686, 2010.

[13] Rong Zhang and Alexander I Rudnicky. A large scale clustering scheme for kernel k-means. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 4, pages 289–292. IEEE, 2002.

[14] Brian Kulis et al. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013.

[15] Mikhail Bilenko, Sugato Basu, and Raymond J Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 11. ACM, 2004.

[16] Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 243–254. SIAM, 2008.

[17] 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.

[18] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

[19] Jake Yang and Xiaojin Zhu. Semi-supervised learning software, 2013. Website containing an assembly of semi-supervised learning software and packages. URL: http://pages.cs.wisc.edu/~jerryzhu/ssl/software.html.

[20] Richard J Bolton and David J Hand. Peer group analysis–local anomaly detection in longitudinal data. Technical report, Technical Report, Department of Mathematics, Imperial College, London, 2001.

[21] Michael A Hayes and Miriam AM Capretz. Contextual anomaly detection framework for big sensor data. *Journal of Big Data*, 2(1):2, 2015.