Data exploration and supervised classification of illicit use 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 supervised machine learning methods. The current misconduct detection system has two stages: first there is a rule-based algorithm which marks access events as suspicious or not. The suspicious events proceed to a second stage where a professional conducts a manual investigation. In the current system, the parameters of the rule-based system are set manually and there is no automatic feedback based on precision of the filtration. Therefore, it is possible that the precision could be improved using supervised machine learning algorithms.

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.

[1] Author's translation. Swedish: Patientdatalagen

[2] Swedish: Stockholms län

[3] Swedish: Vårdgivare

[4] Swedish: Vårdavdelning

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 are separated into different projects. This way, aims can be formulated independently and different approaches can be taken. In this project, the aim is focused on reducing the number of false-positives by training machine learning algorithms in a supervised manner using the labelled data available from the manual investigations. Furthermore, since the data has not been used before in a machine learning context, and since it is desirable for the outcome to be interpretable, a thorough data exploration should be conducted before training the algorithms.

3 Aim and research question

The aim of this degree project is to evaluate machine learning algorithms on their performance with regards to classification of licit and illicit access of patient data in one of Stockholm's county's medical record systems. In order to understand the outcome and diagnose poor results, a thorough data exploitation phase should accompany the training of the algorithms.

The aim can also be formulated as a research question: How can supervised machine learning algorithms be used to classify licit and illicit use of medical records given the available access logs? What difficulties arise from the data set and how can these be addressed? How does these algorithms compare in terms of classification error with the current rule-based system? In terms of computational expenditure?

4 Project evaluation

The implementation of an algorithm which achieves a non-trivial classification error (i.e. better than the ratio of data samples for each class) will be used to evaluate whether supervised learning can be used for this task. If a non-trivial classification error cannot be achieved, the obstructive difficulties should be analysed and explained. An algorithm with non-trivial performance should be compared with the current rule-based system in terms of precision and recall, and the differences be explained.

The explanation and analysis of results should be based on findings from the data exploration. Ideally, it should be understood which data points are more easily classified for the respective algorithm and which features are important for the classification to succeed.

5 Proposed procedure

The project can be divided into three phases: pilot study, data exploration and algorithm evaluation.

5.1 Pilot study

Read literature and make preparations for the rest of the project. Things to read about: **t-SNE**, dimensionality reduction (with **PCA**), how to handle categorical features, how to preprocess log data, supervised machine learning (**neural networks**, **support vector machines**, **random forest**).

5.2 Data exploration

In order for motivated choices to be made regarding which algorithms to test and which features to use, the data must be understood. It is possible that the machine learning algorithms will not be able to make a statistically significant improvement over the existing rule-based system. If this would be the case, knowledge of the structure of the data is key to explaining why. Furthermore, a thorough analysis of the data could prove useful for any future work with this data set.

An important part of this phase is to consider how the data should be represented, i.e. how to preprocess the data and which features to use. A good baseline is to use the features currently employed in the rule-based system. Note, however, that in the current system, the data for each care-unit is treated separately. It is an open question if better result could be achieved by treating the data as one set instead of separating it. Evidence supporting either approach would be an interesting finding of the data exploration phase.

There are several tools available for data exploration. Visualisation is helpful for building an intuition over the structure of the data. Histograms and bar plots are useful when using one feature at a time. A visualisation method which make use of all features is **t-SNE**[5] [1]. Another possibility is to do a dimensionality reduction either using unsupervised methods, e.g. **principal component analysis (PCA)** or **canonical component analysis (CCA)** [2, Chapter 12], or supervised methods, e.g. **partial least squares (PLS)** [3]. Note that many of the features are categorical, so the methods may have to be adjusted to account for this.

The data exploration phase should lead to a decision in these questions:

- How much of the data should be used? Should only the samples which have been manually labelled be used, or should the data filtered away by the current rule-based system be included?
- Is the data unbalanced with respect to the prediction labels? If so, which method should be used to alleviate this problem?
- Which feature representation should be used? Is there any feature which stands out as being more correlated with the class labels than the rest?
- Should the data of each care-unit be treated separately? If not, should a feature be introduced indicating the specific care-unit?
- Are there any machine learning algorithms which appears particularly suitable for the data? Are there any which seem unsuitable?

5.3 Algorithm evaluation

Depending on the results of the data exploration phase, 2-3 different machine learning algorithms should be chosen and their performance compared. The performance of the current rule-based system with manually set parameters should be used as a baseline.

Some suggestions of models to consider are:

- **Support vector machines (SVM)** [4, Chapter 7] if the number of labelled data samples is less than 10^4.
- **Neural networks (NNs)** [4, Chapter 5] if more than 10^5 labelled samples are available.
- **Random forest (RF)** [5]

[5] https://lvdmaaten.github.io/tsne/

- The current rule-based model but learning the parameters instead of using manually set values. In this case, the rule-based model can be considered a **decision tree**.

References

[1] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[2] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[3] Saikat Maitra and Jun Yan. Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Applying multivariate statistical models*, 79: 79–90, 2008.

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

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.