A Machine Learning Approach Towards Early Detection of Frequent Health Care Users

Antti Airola[1], Tapio Pahikkala[1], Heljä Lundgrén-Laine[2,3], Anne Santalahti[4], Päivi Rautava[5], Sanna Salanterä[2,3], Tapio Salakoski[1]

[1]Department of Information Technology, University of Turku, Finland
[2]Department of Nursing Science, University of Turku, Finland
[3]The Hospital District of Southwest Finland
[4]Turku City Healthcare Center, Turku, Finland
[5] Turku University Hospital Research Centre, Finland
{antti.airola,tapio.pahikkala,helja.lundgren-laine,paivi.rautava,tapio.salakoski,sanna.salantera}@utu.fi
anne.santalahti@turku.fi

Abstract. In primary health care, a small number of frequent users incur a large portion of the total health care expenditures. In this work, we study whether it is possible to recognize these frequent users early on, through the application of machine learning based text mining techniques on clinical notes. We implement our study on a data set of 147 Finnish primary health care users, using a regularized least-squares based ranking method. The method achieves a ranking accuracy of 0.68 when making predictions based on the recorded text and observed visitation frequency after 20 visitations by a patient, demonstrating that it is possible to make useful predictions about the future rate of visitations.

Keywords: Clinical Text Mining, Machine Learning, Regularized Least-Squares

1 Introduction

Rising health care costs, overburdened emergency rooms, coverage and adequate resources of health care are currently causing various problems in many countries. One reason that has been focused on is those individuals who repeatedly visit health care clinics and excessively utilize different hospital services. Previous studies have confirmed that this small number of frequent users incur a large portion of health care expenditures [7]. In a Canadian study [4] frequent users of health care were defined with population-based data. Those individuals that visited emergency department more than 7 times were counted as frequent users composing almost 10% of all visits. Patients with 18 visits or more were considered as highly frequent users composing nearly 4% of these visits. Locally the

The 4th International Louhi Workshop on Health Document Text Mining and Information Analysis (Louhi 2013), edited by Hanna Suominen

number of the visits can represent even higher figures up to one fifth of all hospital emergency visits [5]. Frequent users represent a heterogeneous group and explicit common characteristics are not found. However, most of these patients have serious health needs related to chronic diseases, mental health problems or drugs and alcohol abuse [2, 4]. According to the literature review increased risk for the frequent use of health care services seems to be with patients aged 25 to 44 and older than 65 years [7]. It has been suggested, that methods which are able to extract epidemiologic patterns and comorbid conditions of frequent users or groups at risk for frequent use of health care services should be developed [10].

In recent years, machine learning based text mining techniques have shown their applicability for the analysis of unstructured clinical text in various problem settings. Examples include automated systems for diagnosis assignment [12], detection of smoking status [3], quality of life prediction [9], and topic segmentation [6]. In this work, we study whether such methods can be developed for early recognition of frequent users of primary health care services. Internationally, the development of models that would allow predicting the time spent by a patient in a hospital in the future from historical data has been spurred by the ongoing Heritage Health Prize competition[1]. Potentially, the ability to recognize these people early in the process would allow early interventions, leading to better treatment outcomes and reduced costs.

We implement a machine learning method on a data set consisting of visitation dates and corresponding electronic notes recorded for 147 Finnish public health care users over a 14 year period. The approach is based on regularized least-squares based ranking [8], combined with linguistic pre-processing, and the results are evaluated using the nested cross-validation approach [11]. The results demonstrate that already after a small number of visitations it is possible to make reasonable predictions about which patients are at most risk of becoming frequent users, though more data may be necessary for developing systems reliable enough for real-world application.

2 Materials and Methods

The data was collected from 9 public health care centers in a large Finnish city. The recording of patient notes in electronic patient records began in 1998, and thus this is the starting point for our study. The data consists of manually anonymized written documentation for 147 Finnish public health care users recorded between years 1998 and 2011. The number of visits for a given patient during this period varies from around 20 to more than 400. Ethical evaluation of the study was conducted by the ethical committee of the hospital before the study protocol was approved.

The question of interest is, can we after a limited number of visits predict, which patients are at most risk of becoming chronic users with a large number of visitations in the future. We cast this problem as a ranking task, where

[1] `http://www.heritagehealthprize.com`

the aim is to be able to order patients, so that the patients with the largest number of expected visitations would be at the top of the ranking, whereas the patients with the lowest number of visits would be at the bottom. We measure the performance of the trained models using the *ranking accuracy*, which can be considered a generalization of the area under ROC curve measure often used in binary classification tasks (see, e.g., [8] for definition). A ranking accuracy of 1 corresponds to a perfect model, and 0.5 corresponds to a random predictor.

For making predictions, we have access to two sources of information. First, we can consider the frequency at which the patients have been visiting during the period when training data has been gathered. A simple baseline rule for making predictions is that patients who visit most often during the data gathering period would continue to do so in the future. As a second source of information, we have access to the text recorded for each visitation.

We represent the textual data using the standard bag of words-representation. Here, the dimensionality of the data is the same as the number of unique words, with each feature recording the number of times a certain word appears in the documentation of a given patient. Due to the highly inflective nature of the Finnish language, we pre-process the data using the FinTWOL tool, which reduces the words to their base forms. Further, we remove all the features that appear less than 10 times in the data. After these pre-processing steps, we are left with a set of 7243 unique terms.

We generate the features as follows. Let k denote the number of visits used to predict the amount of future visits. We let k range from 2 to 20 in our experiments, and use the frequency of visitation during a 5-year period following the last visit as the label to be predicted. We use the text from first k visitations transforming it into the bag of words-representation, and normalize the resulting feature vector to unit length. As additional feature, we consider the average time length measured in days between two consecutive visits for the patient.

We use a linear model of the type

$$f_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i,$$

where $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$ is the feature vector representation of a data point for which the prediction is to be made, and $\mathbf{w} = (w_1, \ldots, w_d)^{\mathrm{T}}$ is a vector representation of the linear model. We train the model using the ranking regularized least-squares (RankRLS) method [8], which is trained by solving the following optimization problem:

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \sum_{i,j=1}^{n} \left(f_{\mathbf{w}}(\mathbf{x}^i) - f_{\mathbf{w}}(\mathbf{x}^j) - y_i + y_j \right)^2 + \lambda \sum_{i=1}^{d} w_i^2 \right\},$$

where $\mathbf{x}^i$ are the feature vectors of the n training data points and y_i their labels. The first term in the objective function measures the fit of the model to the training data, the second term the complexity of the prediction function, and

λ is a parameter. The fit is measured with the pairwise squared loss, a convex approximation of the ranking error. The minimizer of the objective function can be found by solving the corresponding system of linear equations.

A central challenge in the evaluation is, how to do parameter selection for λ, and at the same time obtain a reliable estimate of the predictive performance. Due to the large dimensionality of the feature space using training set accuracy is unreliable, while at the same time we cannot afford a separate test set due to the small sample size. We use nested cross-validation, where an inner cross-validation loop is used for parameter selection, and an outer one for performance estimation [11]. As the outer cross-validation loop, we use 10-times repeated 10-fold cross-validation, and as the inner loop we use leave-pair-out cross-validation [1]. All the ranking accuracies are computed using the averaging method, where the performance is computed for each fold separately and then averaged.

3 Experimental Results

The experimental results are presented in Figure 1. We plot on x-axis the number of visits used for generating the features, and on y-axis the ranking accuracy. We compare three methods. The first makes predictions based on only the average time between previous visitations, the second only based on text, and the third one combines these two information sources.

The average time between visitations turns out to be a powerful feature, as using it alone one can make more accurate predictions, than based on the text. Intuitively this is not very surprising, one can expect that the visitation rate in the past would to a large degree correlate with that in the future. While the text does not alone work as well, it does represent an alternative source of information about the properties of the patient not captured by the one single numerical feature. Indeed, the combination of using text and the numerical feature for prediction outperforms both approaches. The larger the number of visits used for generating the features, the better the predictive performance becomes, peaking at 0.68 ranking accuracy. This resulting performance is not very high, but nevertheless demonstrates that there is useful information present in the health records for the purpose of estimating future visitation frequencies.

Next, we explored which of the textual features are the most informative ones, considering the data generated for value $k = 20$. For each feature, we computed the ranking accuracy for using it alone as a predictor for the number of future visitations (accuracies below 0.5 were transformed by $1 - \mathrm{ACC}$). We present the 10 highest ranked features in Table 1 in order to provide some further insight into the data. However, it should be noted that due to the very large number of features, some of them can appear informative simply due to random chance.

Of the 10 highest ranked words, 7 describe laboratory tests and 2 are abbreviations for a hospital. These words frequently appear together, describing the event that a patient has been at laboratory examinations at the hospital. The last word *kertaa* (*times*) appears usually in sentences describing medication that has been assigned to a patient (i.e. *MEDICATION 2 times a day*). Thus, the

Table 1. 10 highest ranked features

Word	Meaning	R. Acc
TKS	Central hospital	0.66
B-HB	laboratory test	0.65
TYKS	University Hospital	0.64
kertaa	times	0.64
fB-Leuk	laboratory test	0.64
B-HKR	laboratory test	0.64
B-Trom	laboratory test	0.64
B-Eryt	laboratory test	0.64
B-PVK	laboratory test	0.63
E-CV	laboratory test	0.63

most obvious signs regarding the risk of becoming a chronic patient that can be recognized in the data using a simple univariate statistic appear to be being assigned to laboratory tests, or being assigned daily medication.

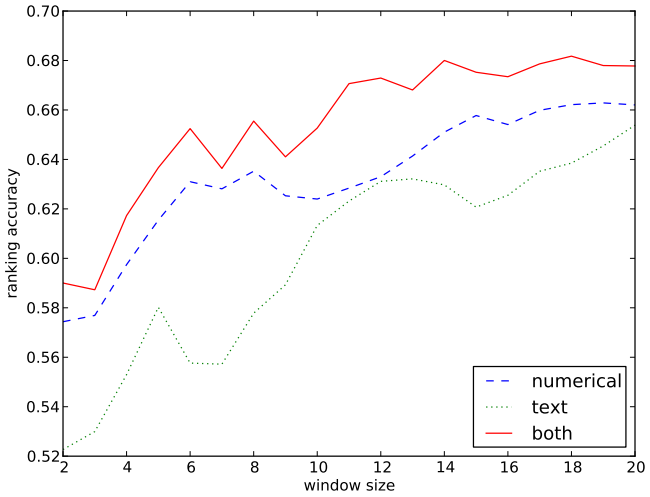

Fig. 1. Ranking accuracies as a function of the number of visits used for training

4 Conclusion

Our results provide a proof-of-concept demonstration about predicting the rate of future visitations for primary health care users, based on previous visitation

frequencies and clinical notes. In order to develop a more accurate system, it would be beneficial to gather more data, as well as additional information beyond free-form textual notes about the patients.

Acknowledgements

This work has been supported by the Academy of Finland. We would like to thank Lingsoft Ltd for collaboration.

References

1. Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., Salakoski, T.: An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. Computational Statistics & Data Analysis 55(4), 1828–1844 (2011)
2. Byrne, M., Murphy, A., Plunkett, P., McGee, H., Murray, A., Bury, G.: Frequent attenders to an emergency department: a study of primary health care use, medical profile, and psychosocial characteristics. Annals of Emergency Medicine 41, 309–318 (2003)
3. Clark, C., Good, K., Jezierny, L., Macpherson, M., Wilson, B., Chajewska, U.: Identifying smokers with a medical extraction system. Journal of the American Medical Informatics Association 15(1), 36–39 (2008)
4. Doupe, M., Palatnick, W., Day, S., Chateau, D., Soodeen, R.A., Burchill, C., Derksen, S.: Frequent users of emergency departments: Developing standard definitions and defining prominent risk factors. Annals of Emergency Medicine 60(1), 24–32 (2012)
5. Fuda, K., Immekus, R.: Frequent users of massachusetts emergency departments: a statewide analysis. Annals of Emergency Medicine 48, 9–16 (2006)
6. Ginter, F., Suominen, H., Pyysalo, S., Salakoski, T.: Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. International Journal of Medical Informatics 78(12), e1 – e6 (2009)
7. LaCalle, E., Rabin, E.: Frequent users of emergency departments: the myths, the data, and the policy implications. Annals of Emergency Medicine 56(1), 42–48 (2010)
8. Pahikkala, T., Tsivtsivadze, E., Airola, A., Järvinen, J., Boberg, J.: An efficient algorithm for learning to rank from preference graphs. Machine Learning 75(1), 129–165 (2009)
9. Pakhomov, S., Shah, N., Hanson, P., Balasubramaniam, S., Smith, S.A.: Automatic quality of life prediction using electronic medical records. In: Proceedings of AMIA Annual Symposium. pp. 545–549 (2008)
10. Pines, J., Asplin, B., Kaji, A., Lowe, R., Magid, D., Raven, M., Weber, E., Yealy, D.: Frequent users of emergency department services: gaps in knowledge and a proposed research agenda. Academic Emergency Medicine (6), e64–9 (2011)
11. Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7(91) (2006)
12. Wang, Z., Shah, A.D., Tate, A.R., Denaxas, S., Shawe-Taylor, J., Hemingway, H.: Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS ONE 7(1), e30412 (2012)