MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-1024

Machine Learning Methods for Clinical Forms Analysis in Mental Health

John Strauss [a], Arturo Martinez Peguero[b], Graeme Hirst[b]

[a] *Centre for Addiction and Mental Health, University of Toronto, Toronto, Canada*
[b] *Department of Computer Science, University of Toronto, Toronto, Canada*

Abstract and Objective

In preparation for a clinical information system implementation, the Centre for Addiction and Mental Health (CAMH) Clinical Information Transformation project completed multiple preparation steps. An automated process was desired to supplement the onerous task of manual analysis of clinical forms. We used natural language processing (NLP) and machine learning (ML) methods for a series of 266 separate clinical forms. For the investigation, documents were represented by feature vectors. We used four ML algorithms for our examination of the forms: cluster analysis, k-nearest neigh-bours (kNN), decision trees and support vector machines (SVM). Parameters for each algorithm were optimized. SVM had the best performance with a precision of 64.6%. Though we did not find any method sufficiently accurate for practical use, to our knowledge this approach to forms has not been used previously in mental health.

Keywords: Psychiatry, Natural language processing, Support vector machines, Clinical research informatics

Introduction

The Centre for Addiction and Mental Health (CAMH) is Canada's largest research and teaching centre focused on mental health and addiction, with over 20,000 unique patient visits annually. CAMH is preparing for the implementation of a new clinical information system (CIS). As part of the CAMH CIS readiness project, we had a corpus of 266 clinical forms requiring classification by the type of form. Document classification is a common task in ML, so we hypothesized that ML methods would be a useful comparison to our manual/analog clinical forms analysis.

Methods

All 266 forms existed in, or were converted into .pdf format. The PDFMiner utility was used to convert .pdf documents to plain text to enable manipulation of all text data from the forms' questions. The information retrieval method term frequency-inverse document frequency (TF-IDF) was used to compute the relevance of each word in a form to the form itself. IDF is maximized when document frequency is low -- i.e. When a word is unique to a single document in a corpus. Each document in the corpus was given a feature vector representation depending on the terms its questions contained. Orange is the open source ML tool suite for Python that employs the ML tools that we used: (decision trees[1], cluster analysis, k-nearest neighbours (kNN), and support vector machines (SVM)) via scripts or with a graphical user interface (GUI).

Results

K-means clustering gave the least promising results. Values of *k* between 2 and 25 were used but failed to produce reasonably partitioned clusters[2], with the highest silhouette value equalling 0.21 for *k*=2. With k-NN the highest precisions were achieved with k=11,12,13,14; the maximum precision was 36.7% without scaling of data. Decision tree precision was asymptotic and fell short of 50%. SVM methods produced better results and performance was optimized for the precision to reach 64.6%.

Conclusion

We applied four ML algorithms to text extracted from a corpus of 266 blank clinical mental health forms. Preliminary clustering results projected poor groupings. Scaling, as suggested [3] proved unhelpful for k-NN. SVM performed best of three classification methods, with several accuracies in the low 60% range, in keeping with its being more capable than other classifiers with small amounts of training data[4]. Overall, our novel application of these methods proved not to be of practical use for streamlining clinical forms. Likely, two or three principal limitations impacted the performance. First, the corpus consisted of a small number of brief documents most often with sparse amounts of text. Second the text on a given form does not always reflect the document class. Lastly, the text elements have a significant degree of overlap across forms.

References

[1] C. D. Manning and H. Schutze, Foundations of Statistical Natural Language Processing . Cambridge, MA: MIT Press, 1999.

[2] Kaufman, Leonard and Rousseeuw, Peter J., Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc., 2008.

[3] C. W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," tech. rep., Department of Computer Science, National Taiwan University, 2003.

[4] C. D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval. Cambridge University Press, 1 ed., July 2008

Address for correspondence
John Strauss
Centre for Addiction and Mental Health
1001 Queen St. West, Toronto, ON M6J 1H4
john.strauss@camh.ca