Knowledge Extraction and Outcome Prediction using Medical Notes

Ryan Cobb, Sahil Puri, Daisy Wang RCOBB, SAHIL, DAISYW@CISE.UFL.EDU
Department of Computer & Information Science & Engineering, University of Florida, Gainesville, FL

Tezcan Baslanti, Azra Bihorac TOZRAZGATBASLANTI, ABIHORAC@ANEST.UFL.EDU
Department of Anesthesiology, University of Florida, Gainesville, FL

Abstract

The increasing use of electronic health records (EHR) has allowed for an unprecedented ability to perform analysis on patient data. By training a number of statistical machine learning classifiers over the unstructured text found in admission notes and operating procedures, prediction of a surgical procedure's outcome can be performed. We extend an initial bag-of-words model to a bag-of-concepts model, which uses cTakes and UMLS to extract medical terms and concepts from medical notes. We also extend cTakes to improve the knowledge extraction. Lastly, we propose a knowledge exchange component, which allows physicians to provide feedback on outcome results to further tune the underlying classifier.

1. Introduction

Medical text notes carry invaluable information about the current and previous medical history, current symptoms and severity of condition as well as physicians clinical judgment. Due to the sheer quantity and unstructured form of text data, it is rarely used for analysis or prediction in a clinical setting. Natural language processing techniques have been used with clinical text data to perform tasks, such as, pre-processing, contextual feature detection and code and entity extraction[7][8][14]. However, there are a vast array of richer concepts that are currently untapped within medical text including: symptoms, medications, surgical events, and procedures performed. If these rich concepts can be effectively extracted, like a real physician, a classifier could be trained to predict the outcome of surgeries.

This paper describes a processing pipeline tailored for health records. The pipeline is divided into three primary categories: outcome prediction, feature extraction and knowledge exchange. We begin with a bag-of-words model and refine it into a bag-of-concepts in order to capture these richer concepts present in health records. We extract these concepts using both cTakes, a state-of-the-art medical extraction tool, and our work in progress extensions of context sensitive section identification and ontology collapsing. These extracted features are then used as training samples in supervised learning of an outcome classifier. The final proposed stage then presents the results to the physician, with which they can provide feedback as active learning for the classifier. We show preliminary results for our current system with respect to classification and feature extraction.

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

2. Data Description

Our dataset consists of 1,999 deidentified patient records obtained from University of Florida Health. In Table 1, we describe the 21 columns of discrete data for each patient record. Many of the discrete data columns can be interpreted literally. Also, a brief description about each data attribute is given with the column name.

In addition to the discrete patient data, we have three columns of medical text data in different stages of patient workup. The first note is a History and Physical examination note by admitting physician. The second note is an operative report written by a surgeon performing the procedure summarizing the technique and all events during surgery. Lastly, discharge notes summarize the events of the hospitalization and provides plan at the time of discharge from the hospital.

3. System Overview

Our system can be divided into three logical subgroups for processing of EHR.

- Outcome Prediction:
  This phase is central to the system and it directly interacts with the other two phases. It involves training a classier for predicting the outcome of a surgery in terms of probability for mortality and

Column	Description
ID	Unique identifier
Admission Type	Admission type (Emergency admission, routine elective admission)
Admission Date	Date of admission
Discharge Date	Date of discharge
Mortality	Hospital Mortality (Yes, No)
RIFLE-AKI	Acute Kidney Injury using RIFLE (Risk, Injury, Failure, Loss, and End-stage Kidney) consensus definition (Yes, No)
Gender	Patient gender
Race	Patient race
Age	Patient age (in years)
Zip	Patient home zip code
Cci	Charlson comorbidity index score
Dx1	ICD-9-CM code for primary diagnosis
Pr1	ICD-9-CM code for primary procedure
Pr1 gr	Two digit grouping of primary procedure group
Pr1 day	Day of primary procedure relative to admission day
Service	Surgery type
Admitting Doctor	Admitting doctor
Attending Doctor	Attending Doctor
H&P notes	History and physical notes
Op notes	Operation notes by doctor
Dc notes	Discharge notes

Table 1. Data description

postoperative complications. The predictor takes discrete data and concepts from the knowledge extraction phase as input. The output of the predictor will collaboratively work with the knowledge exchange phase.

- Knowledge Extraction:

  We use knowledge extraction as a preprocessing step to vectorize EHR for use in the other stages. This phase extracts medical concepts in broad categories of Problems, Tests and Treatments. Additionally, it enriches extracted concepts with the contextual attributes of negation, temporality and experiencer.

- Knowledge Exchange:

  The final subsystem is a post processing step taken to incrementally increase classification accuracy over time by soliciting feedback from expert users. We describe two forms of knowledge exchange received from users during classifier training and real-world utilization.

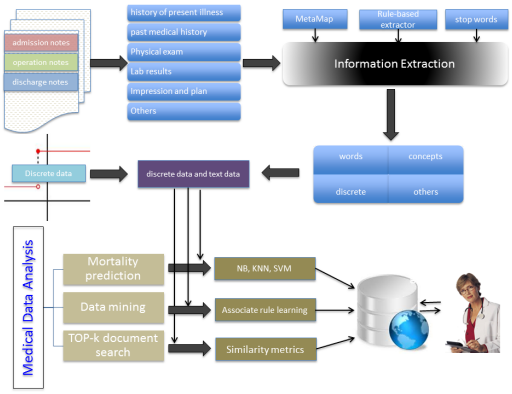

Figure 1. System Overview

4. Outcome Prediction

The goal of text analysis on EHR data is to augment information input value beyond discrete data to improve prediction of outcomes during patient care. We leverage not only the large corpus of discrete and text data, but also the use of expert knowledge in fine tuning machine learning algorithms.

4.1. Classifiers

We compared five different supervised classification algorithms based upon their classification performance and ease of retraining during expert knowledge exchange. The scikit python library[10] was used to compare Neural Networks, K-Nearest Neighbors, SVM, Nearest Centroid and Naive Bayes (Multinomial and Bernoulli) classifiers. We used the F1-score and classification accuracy as quantitative measures of classifier performance and we qualitatively determined the feasibility of tuning the underlying model. Using these metrics, we were able to rank the classifiers for our prediction task.

Due to the need for knowledge exchange between the classifier and expert, we heavily weighted this consideration for each classifier. Active learning for classifiers is typically performed in one of two forms: new sample instances for training or direct manipulation of the classifier weights. The underlying model for two of our classifiers, SVM and Neural Networks, create complex hyperplanes in large multi-dimensional spaces and prove difficult to visualize when tuning through direct manipulation. The fundamental naive independence assumption of the Naive Bayes classifier facilitates an expert to independently assess the concept's relevance to an outcome of patient care. This can be contrasted to the Nearest Centroid classifier where relevance feedback is the comparison of two documents (pa-

tient EHR) and how relevant they are to each other. In the domain of document search, the Nearest Centroid classifier has been effectively used by search engines[4], the comparison process is too time consuming for an expert compared to the concept relevance feedback of Naive Bayes. Thus, we have chosen to use the Naive Bayes classifier as our baseline classifier. As future work, we are planning to extend the Naive Bayes model to incorporate dependencies between features.

4.2. Features

Bag-of-words model has been extensively used in document classification[6]. In the bag-of-words model, text is represented as an unordered collection of words, disregarding grammar and word order. The occurrence of or frequency of each word is used as a feature for training a classifier. However, EHRs contain variable-length terminologies and proper extraction of these terminologies is crucial for accurate outcome prediction. A single medical terminology may contain a group of one or more words. For example, "obesity" has only one word and "alcoholic intoxication" contains 2 words. To address this aspect of EHR, we use the "bag-of-concepts" model. In the bag-of-concepts model, the adjacent words are combined into a single medical concept using medical dictionaries. We use the Unified Medical Language System (UMLS)[2] and cTakes[12] to extract variable-length concepts from EHR.

In addition to the textual concepts found in the discrete data and notes, EHR contains continuous numeric data such as age. The textual concepts easily fit into both the bag-of-words and the bag-of-concepts as categorical features. However, the continuous data cannot be easily added because categorical data is expected in a bag-of-words model. We naively discretize the continuous data into sets of fixed interval categorical bins. We plan to investigate the Non-Disjoint Discretization method[15] as a means to more accurately represent the true categorization of continuous values.

4.3. Initial Results

We evaluated each of the classifiers with the following feature models: bag-of-words and bag-of-concepts. The initial results showed that bag-of-words out performed bag-of-concepts in all of the evaluated classifiers. We observed that the dimensionality reduced from the bag-of-words to bag-of-concepts model, but the instance count of all features in bag-of-concepts was also reduced dramatically. As a result, the feature space for bag-of-concepts was much sparser than the bag-of-words model. With this in mind, we have proposed several modifications in the following section to resolve the accuracy issues of the bag-of-concepts model.

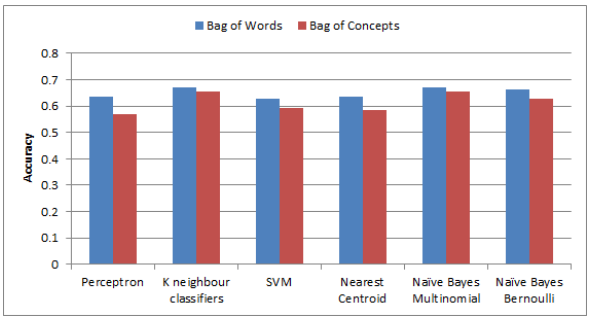

Figure 2. Classifier Accuracy with Bag-of-Words vs Bag-of-Concepts

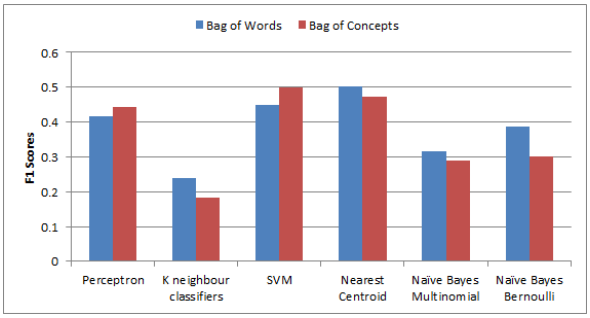

Figure 3. Classifier F1 Score with Bag-of-Words vs Bag-of-Concepts

5. Extracting features from Medical Notes

Effective feature extraction is crucial in learning accurate classifiers. Given our initial classifier results, we are pursuing higher quality methods to reduce the curse of dimensionality that arises from medical concepts. Each of the following methods we are proposing is intended to either reduce dimensionality or increase feature strength. At present, we have restricted the feature extraction to be only from History and Physical notes, but plan to incorporate data from operating and discharge notes as well.

5.1. Discover numeric values features from the text

Medical notes contain several important discrete features that can be extracted, such as vital signs, lab results and medication dosage. Typical vital signs reported for a patient are features such as blood pressure, pulse, respiration, oxygen saturation, temperature and other findings on physical examination. Lab results are also an important indicator of the physical status of the patient and have been shown to be key indicators of complications[1]. In order to extract these features, we use cTakes' specialized modules, lab results classifier and drug named-entity recognition, for extraction of these numeric values. We will extend the cTakes classifier toolkit to also include classification of vital signs and physical indicators. As mentioned in Section 4.2, these

numeric values will be discretized and appended to their corresponding medical concept to create a unique feature associating the concept to its discretized value.

5.2. Section Identification

Commonly, admission notes are divided into several sections. They are chief complaints, history of present illness, past medical history, family and social history, medication list, physical exam, lab result and assessment and plan. Given that conceptually similar medical concepts may be present in different sections, but they can be semantically different depending on the implicit meaning of the section. We wish to capture these feature qualities by developing a rule-based system to partition the document into their corresponding sections. These resulting partitions are then individually passed into cTakes for feature extraction.

Of the sections within EHR, three of them were explicitly shortlisted by domain experts as containing the most pertinent information for outcome prediction. These sections were History of Present Illness, Physical Examination and Assessment and Plan. The history of present illness focuses on *symptoms* observed by the patient. The physical exam provides standard observations about the patient's status by physician. Lastly, the assessment and plan contains *signs* and the doctor's subjective summary of a patient. In future work, we will focus on extraction of the doctor's sentiment as means of detecting the severity of a patient's condition[3][9].

5.3. Sparseness of Features

Given the sparsity of samples for any given concept, we are proposing to map extracted features onto a medical concept ontology and then collapse sparse nodes into their parent concept to increase the sample instances of a concept. Specifically, we use the UMLS and ICD-9 ontologies to deterministically collapse the concepts based upon sample density until we hit a specified sample threshold. The method of collapsing concepts helps ensure a statistically significant number of samples for training the classifier.

6. Knowledge Exchange

Due to the inconsistencies that may arise from extraction errors or sparse data, we propose to exchange knowledge between the expert and the classifier. Presenting classifier extracted knowledge to the expert can highlight previously ignored patient details. Likewise, by soliciting knowledge from experts, the model's feature quality and classification accuracy can be boosted. This feedback will then be integrated as new training instances for active learning of the classifier[5][11][13].

We ran a pilot study of our knowledge exchange interface

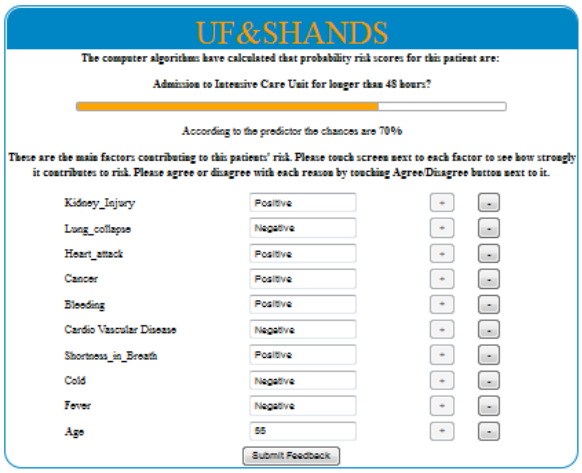

Figure 4. Pilot Test Knowledge Exchange Interface

with two physicians evaluating 50 patients. In this study, we tested the physician's acceptance of the important features and classifier results. In order to assess the impact of our system on the physician's opinion, we divided the study into three steps. The system provided summarized details of the patient's condition and solicited for the physician's assessment of a patient spending more than 48 hours in the ICU. In the next step, the system displayed the classifier's assessment and the top 10 contributing factors for increased risk. The physician evaluated each justifying factor through approval or disapproval as shown in Figure 4. Finally, the physician was allowed to revise their original assessment based on the classifier's features. We found the physician's revised assessments were on average 13% more accurate after knowledge exchange (CI 95%, p-value 0.02). On average, 78.2% of the features were approved of by the physicians.

7. Conclusion

We have presented the problem of patient outcome prediction and risk stratification. In the medical domain, a physicians time is a precious and expensive resource and any task that consumes time is an expensive proposition. We have proposed a system that aims to reduce the amount of time a physician spends gathering information and formulating a decision regarding risk for adverse outcomes using EHR. The proposed system will automatically classify and quantify patient's personal risk for mortality and complications. The result is displayed as summarized content contributing to the classification allowing the doctor to expeditiously assess the situation, provide feedback and augment their own clinical judgment and decision making.

References

[1] Azra Bihorac, Matthew J Delano, Jesse D Schold, M Cecilia Lopez, Avery B Nathens, Ronald V Maier, A Joseph Layon, Henry V Baker, and Lyle L Moldawer. Incidence, clinical predictors, genomics, and outcome of acute kidney injury among trauma patients. *Annals of surgery*, 252(1):158, 2010.

[2] Olivier Bodenreider. The unified medical language system (umls): Integrating biomedical terminology, 2004.

[3] Ann Devitt and Khurshid Ahmad. Sentiment analysis and the use of extrinsic datasets in evaluation. In *Proc. of the 6th Intl. Conf on Language Resources and Evaluation*, 2008.

[4] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[5] Andrew McCallum, Gideon Mann, and Gregory Druck. Generalized expectation criteria. *Computer science technical note, University of Massachusetts, Amherst, MA*, 2007.

[6] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[7] Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, John F Hurdle, et al. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, 35:128–44, 2008.

[8] Mei-Sing Ong, Farah Magrabi, and Enrico Coiera. Automated identification of extreme-risk events in clinical incident reports. *Journal of the American Medical Informatics Association*, 19(e1):e110–e118, 2012.

[9] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, 2001.

[12] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[13] Simon Tong. *Active learning: theory and applications*. PhD thesis, Citeseer, 2001.

[14] Xue T Wee, Yvonne Koh, and Chun W Yap. Automated systems to identify relevant documents in product risk management. *BMC Medical Informatics and Decision Making*, 12(1):13, 2012.

[15] Ying Yang and Geoffrey I Webb. Non-disjoint discretization for naive-bayes classifiers. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 666–673, 2002.