PANDA: Predictive Analysis with iNtegration of Doctoral Assessments

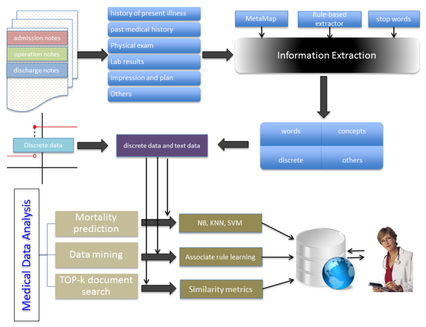
Ryan Cobb[§], Sahil Puri[§]
[§]CISE, University of Florida
{rcobb, sahil}@cise.ufl.edu

UF UNIVERSITY of FLORIDA
The Foundation for The Gator Nation

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[2]. We extend an initial bag-of-words model to a bag-of-concepts model, which uses cTakes[1] 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.

System Overview



Challenges

- Expert knowledge integration into classifier.
- Higher quality feature extraction.
- Representation of feature dependencies.

Acknowledgements

- Tezcan Ozrazgat Baslanti, Ph.D.
- Azra Bihorac, MD MS
- Kun Li
- Sean Goldberg

Feature Extraction

Shallow parser output:

```
Fx  of  obesity  but  no  fx  of  coronary  artery  diseases  .
NP  PP  NP       PP   NP  NP  PP  NP
```

Named Entity Recognition – 5 Named Entities found:

```
Fx of obesity but no fx of coronary artery diseases .
     obesity  (type=diseases/disorders, UMLS CUI=C0028754, SNOMED-CT codes=308124008 and 5476005)
                        coronary artery diseases (type=diseases/disorders, CUI=C0010054)
                        coronary artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
                                artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
                                diseases (type=diseases/disorders, CUI = C0010054)
```

Status and Negation attributes assigned to Named Entities:

```
Fx of obesity but no fx of coronary artery diseases .
     obesity (status = family_history_of; negation = not_negated)
                        coronary artery diseases (status = family_history_of, negation = is_negated)
```

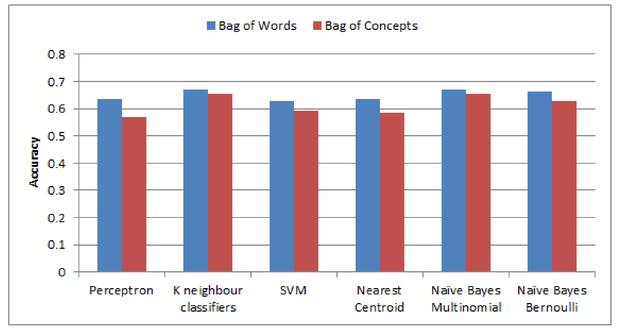
We use knowledge extraction as a preprocessing step to vectorize EHR for use in the other stages. This phase extracts medical concepts in broad categories of Problems, Tests and Treatments using the cTakes system. The following feature extractions enhancements are a work-in-progress:

- **Numeric Test Results**: Extracted numeric values are correlated with their test concept and discretized to fit into bag-of-concepts model.
- **Section Selectivity**: History of Present Illness, Physical Examination and Assessment and Plan are extracted based on their medical importance.
- **Sparseness of Features**: Ontology collapsing on the ICD-9 tree to increase concept frequency for better training of classifiers.

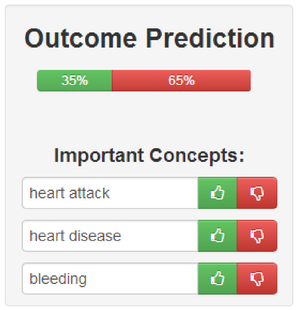
Outcome Prediction

Using supervised learning, we take the results of the Feature Extraction phase and the known outcome to train a set of classifiers. We compared two models of feature representation: bag-of-words and bag-of-concepts. The metrics of evaluation were accuracy and transparency of the classifier for human comprehension.

- **Bag-of-Words**: Every word is considered as a feature.
- **Bag-of-Concepts**: Every medical concept is a feature.
- **Classifiers**: We shortlisted Naive Bayes and Nearest Centroid based on the metrics above. Accuracy results are in the figure to the right.



Knowledge Exchange



Due to the inconsistencies that may arise from extraction errors or sparse data, we propose to incorporate expert knowledge as a method to improve the model further. By soliciting knowledge from experts, feature quality and classification accuracy can be boosted. To this end, these expert-in-the-loop interactions will manifest themselves in two phases: model building and testing.

We present the top-k features to the medical experts who give feedback based upon their real-world experience. The feedback is integrated as new training instances for active learning of the classifier.

References

[1] Savova, Guergana K., et al. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES)." JAMIA 17.5 (2010): 507-513.
[2] Bihorac, Azra, et al. "Incidence, clinical predictors, genomics, and outcome of acute kidney injury among trauma patients." Annals of surgery 252.1 (2010): 158.