Evaluating classification power of linked admission data sources with text mining

Simon KOCBEK[a,b,1], Lawrence CAVEDON[a], David MARTINEZ[b,c], Christopher BAIN[d,e], Chris MAC MANUS[d], Gholamreza HAFFARI[e], Ingrid ZUKERMAN[e], Karin VERSPOOR[b]

[a]*Computer Science & Info Tech, RMIT University, Melbourne*
[b]*Dept of Computing and Information Systems, University of Melbourne, Melbourne*
[c]*MedWhat.com, San Francisco*
[d]*Health Informatics Department, Alfred Hospital, Melbourne*
[e]*Faculty of Information Technology, Monash University, Melbourne*

Abstract. Lung cancer is a leading cause of death in developed countries. This paper presents a text mining system using Support Vector Machines for detecting lung cancer admissions. Performance of the system using different clinical data sources is evaluated. We use radiology reports as an initial data source and add other sources, such as pathology reports, patient demographic information and hospital admission information. Results show that mining over linked data sources significantly improves classification performance with a maximum F-Score improvement of 0.057.

Keywords. Text mining, natural language processing, lung cancer, linked hospital data

Introduction

Text and data mining are proving to be increasingly important and powerful techniques for extracting information and insights from Health and Hospital Information Systems [1-5]. Mining hospital data holds the potential for new discoveries as well as improved efficiencies and communication within hospital systems. Much valuable information in hospital records is represented in free text format, e.g., radiology and pathology reports, requiring the application of Text Mining (TM) and Natural Language Processing (NLP) techniques.

Most previous clinical text mining applications have made use of a single textual data source, e.g., radiology reports, in order to identify or mine information related to a single condition (e.g., [1,2]). However, the increase in data linkage (i.e., multiple data sources being linked by patient id) in Hospital Information Systems is creating opportunities for more powerful and accurate text mining techniques that combine insights from multiple data sources [6].

In this paper, we describe performance of text mining in the context of the challenge of identifying patients admitted to a hospital for treatment for lung cancer. Lung cancer is a leading cause of death in developed countries, and automatically

[1] Corresponding Author.

mapping patient admissions to ICD (International Code of Diseases) directly from hospital records is a precursor to automated ICD-coding, a massively time-consuming manual process at the core of the procedure followed to fund hospitals.

The focus of this paper is to evaluate the value of data linkage and investigate the source of value within different hospital data sources. In particular, we consider a large collection of radiology and pathology reports, along with associated metadata sources, and build classifiers for each type of data source, as well as their combination. Our results confirm that, as might be expected, jointly mining multiple linked data sources improves text classification performance. Analysis also identifies which information source is most valuable for mining for the specified disease, although we expect this to vary with different diseases.

1. Related work

A substantial amount of relevant disease information exists in various types of medical records. Much of this information is in the form of free text; hence text mining represents a promising strategy for building machine learning classifiers that take advantage of the richness of such records. Both radiology and pathology reports have been studied as a source of specific clinical information in previous text mining studies. A pathology report describes the results of examining cells and tissues under a microscope after a biopsy or surgery. A radiology report represents a specialist's interpretation of images related to a patient's signs and symptoms.

Hripscak et al. [1] used NLP techniques to evaluate the automatic coding of 889,921 chest radiology reports. Nguyen et al. [2] performed classification of lung cancer stages from pathology reports. In their follow-up work [3], a rule-based system was used to classify cancer-notifiable pathology reports from a small corpus (approx. 500 reports), obtaining very high sensitivity, specificity and Positive Predictive Value (PPV). Pathology reports have also been analysed to extract breast cancer characteristics into a knowledge model [4] and to identify relevant named entities [5].

In previous work [7] we built a system for detecting lung cancer admissions based on radiology reports linked to patient metadata for the financial years 2012-2013 and 2013-2014. A similar approach is adopted in this paper, where we use TM techniques to extract useful information about lung cancer. We extend the prior work by exploring the impact of incorporating two additional data sources: pathology reports and radiology questions (i.e., the purpose stated by the clinician for requesting a scan). We also measure statistical significance of classification performance using the different data sources. Note that the goal of this paper is not to achieve better classification performance than previous systems, but to achieve comparable performance and explore the value of various data sources in mining information related to a specified question.

2. Methods

2.1. Data source

The data for this study was extracted from the Alfred Health Informatics Platform, called REASON [8], which provides a single data warehouse view of multiple data

sources within the Alfred Health system, linked by unique anonymised patient id. Data for the current study was extracted from REASON under ethics approval from the Alfred Health Human Research Ethics Committee, in the form of a de-identified set. A high-level architecture of the REASON platform is shown in Figure 1. Table 1 provides an overview of some of the key record types (and number of records) in the platform relevant to our current task, though it is not a complete listing.

For the purpose of this study, we extracted textual form of radiology and pathology reports for the financial years 2012-2013 and 2013-2014. Each report was assigned an admission identifier, which is in turn linked to patient metadata. The following metadata associated with each admission were extracted: patient's demographic data (gender, age, ethnic origin, country, language, marital status, religion, and death date) and hospital-related admission data (hospital code, admission date and time, discharge date and time, length of stay, reason for the admission, admission unit, discharge unit, admission type, source, destination and criteria). Radiology reports were also associated with radiology *questions*, i.e., a short description of the reason given by the clinician for requesting the scan. The initial number of admission records used in this study was as follows: 40,800 radiology reports; 20,872 pathology reports; and 121,700 metadata entries.

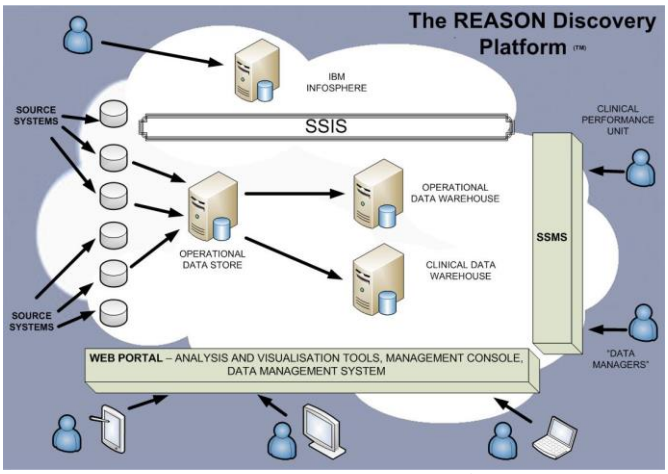

Figure 1. A high level architectural view of REASON.

Table 1. Example of numbers of records by type in REASON.

Data	Record numbers
Admissions	881,653
Emergency Encounter	912,931
Pathology Results- Atomic	43,606,065
Pathology Results- Textual	667,303
Patients	1,884,527
Pharmacy Drug Dispense Transactions	4,131,227
Radiology Reports	756,164
Radiology Test Orders	792,312
Surgeries Performed	158,853

2.2. Gold Standard data set

Each admission is associated with a set of ICD-10 codes, which are annotated in the admission record by an internal clinical coder for reporting purposes. These are used in our study as ground truth to build the gold standard data set. The ICD codes are ignored when testing the classifiers – i.e., the classification task consists of identifying those records which contain the ICD code of interest in the gold standard data set.

To identify positive lung cancer cases we used the ICD-10 code *C34.**: *Malignant neoplasm of bronchus and lung.* In our dataset, only 496 out of 40,800 admissions with radiology reports were positive for lung cancer. The highly skewed nature of the data poses a specific challenge to automated machine learning approaches, which generally perform better over balanced class distributions. To address this problem, we performed subsampling, randomly selecting a subset of negative admissions to balance the datasets. Other, more time complex methods (such as oversampling [9]) could have been used; however, due to time constraints and the high number of experiments to be run, these methods were not appropriate for this work. The final gold standard dataset therefore contained 992 admissions. All admissions contained radiology report and radiology question, 833 admissions also contained metadata, and 518 admissions also contained pathology report.

2.3. Data representation

Machine learning algorithms require a representation of relevant features of each data point that can be used to build a predictive classifier. The feature representation we adopted for our task combines characteristics obtained from text reports, along with the patient and hospital metadata linked to each admission.

Text in radiology reports, radiology questions and pathology reports was processed with the MetaMap tool [10] from the US National Library of Medicine. MetaMap is a program that identifies and normalises biomedical terminology from the Unified Medical Language System (UMLS) Metathesaurus in biomedical text. Below is a short sample of MetaMap-annotated phrases from the sentence "*replaced with a right frontal approach*".

Meta Mapping (701):
748 C0559956: Replaced (Replacement) [Functional Concept]
748 C0205090: Right [Spatial Concept]
778 C2316681: Frontal approach [Functional Concept]

We employed the NegEx module to identify the polarity (negative or positive, e.g., "*Non contrast in the brain*") of phrases. NegEx is a simple algorithm included in MetaMap that implements several regular expressions that indicate negation, filters out sentences containing phrases that falsely appear to be negation phrases, and limits the scope of the negation phrases [11].

We collected phrases mapped into UMLS concepts for each sentence. Identified phrases were marked with whether the concepts were found in a positive or negative context. Phrases from different reports of the same kind (e.g., radiology reports) belonging to the same admission were merged such that repeating phrase was counted only once. We then built series of feature vectors *r[_q][_p][_m]*. The *r* feature vector represents our baseline and contains a "bag" (i.e., an unordered list) of biomedical phrases from radiology reports only. Other feature vectors add the following optional sources: *q – radiology questions*, *p – pathology reports*, and *m – metadata*.

2.4. Classification and evaluation

We treated ICD-codes as targets for classification. To identify those data sources that contain the most valuable information for identifying lung cancer admissions, a classification framework was built for each feature vector described above.
We used the Weka Toolkit [12] implementation of the Support Vector Machine algorithm, since it has performed robustly in our previous work [7].

Evaluation of TM and NLP systems typically involves the following three metrics: *precision*, *recall* and *F-Score*. Precision of positive/negative class (also called positive/negative predictive value) is the ratio of correctly classified positive/negative values to the number of all instances classified as positive/negative. Recall of positive/negative class is computed as the number of correctly classified instances from the positive/negative class divided by the number of all instances from the positive/negative class; this is also known as *sensitivity*. F-score is the weighted harmonic mean of precision and recall.

We performed 10-fold cross-validation, where we randomly split data into train/test halves 10 times. We measured precision, recall and F-Score for each fold. We calculated statistical significance for F-Score using the Wilcoxon signed-rank test, as recommended in [13].

3. Results

Table 3 shows precision, recall and F-Score measurements for the SVM classifiers built for 8 different combinations of feature vectors. The classifier with the lowest score (r) correctly classified 801 admissions, while the classifier with the highest score (r_q_p_m) correctly classified 915 admissions.

Table 3. Precision, recall and F-Score for classifiers built on 8 different feature vectors.

	r	**r_q**	**r_p**	**r_m**	**r_q_p**	**r_q_m**	**r_p_m**	**r_q_p_m**
Precision	0.875	0.898	0.888	0.902	0.906	0.912	0.920	0.932
Recall	0.873	0.896	0.886	0.901	0.904	0.911	0.917	0.930
F-Score	0.873	0.896	0.886	0.901	0.904	0.911	0.917	0.930

Table 4 shows F-Score differences between pairs of classifiers with different combinations of feature vectors. Column names are initial combinations and row names add information. Boldfaced values represent statistically significant results, and comparisons that are not applicable have no values. For example, the left top cell represents the F-Score difference between the classifier built on phrases from radiology reports and the classifier with added radiology question phrases (r_q).

Table 4. F-Score differences between pairs of classifiers and statistical significance.

	r	**r_q**	**r_p**	**r_m**	**r_q_p**	**r_q_m**	**r_p_m**
+q	**+0.023**		**+0.018**	+0.010			+0.013
+p	+0.013	+0.008		+0.016		**+0.019**	
+m	**+0.028**	+0.015	**+0.031**		**+0.026**		

As can be seen, the enhanced classifier performed significantly better than the baseline system (r), with an F-Score difference of +0.023. Similarly, the top right cell shows

that a classifier which uses all the features (r_q_p_m) performs better than the classifier without radiology question phrases (r_p_m); however, this difference was not statistically significant.

4. Discussion

Our baseline classifier for automatically identifying cases of lung cancer built on only radiology report phrases shows comparable performance to that in our previous work [7] (results are not directly comparable since the two datasets involve different timeframes). Precison, recall and F-Score yield similar results for single feature vector (single column in Table 3), which indicates that our classifiers misclassified similar number of positive and negative examples. Including additional admission data sources improved classification performance. The classifier with the highest performance was built using features from all four data sources. However, statistical tests showed that not all performance increases were significant. An example of a non-significant improvement is combining radiology reports with pathology reports (First column in Table 4, r+p). In contrast, adding metadata or radiology questions to radiology reports significantly improved performance. In addition, these two data sources significantly improved the performance when added to already combined radiology and pathology reports (third column in Table 4). Finally, adding metadata to already combined radiology and pathology reports with radiology questions further improves performance (Column 5 of Table 4). Pathology reports significantly increased performance only when added to the combination of radiology reports, radiology questions, and metadata.

Not unexpectedly, our results indicate that more informed systems can be built by including multiple data sources. Radiology questions and metadata seem to contain crucial information for detecting lung cancer cases, significantly improving performance when added to radiology reports or to the combination of radiology and pathology reports. The reason for lack of statistical significance when adding pathology reports to train the system may be due to a dearth of pathology reports (only 518 of 992 admissions with a radiology report had pathology reports associated with them).

5. Conclusion

We have shown that mining multiple linked data sources improves classification performance of lung cancer ICD-10 codes from textual data, as compared to using a single data source. We expect similar results for other diseases and plan to use different ICD-10 codes as targets for classification in our future work. In addition, we plan to use other techniques to address the problem of highly skewed data sets such as oversampling [9] or cost-sensitive learning [14]. Finally, we plan to use methods for identifying features from specific data sources that most influence classification performance. Our data have a high number of features compared to number of samples, and we expect that some of these features are redundant or irrelevant: we plan to apply feature selection methods [15], which should also shorten model training times on the whole dataset and reduce the potential of over-fitting to the data.

References

[1] G. Hripcsak, J.H. Austin, P.O. Alderson, C. Friedman, Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports, *Radiology*, **224** (2002), pp. 157–163.

[2] A.N. Nguyen, M.J. Lawley, D.P. Hansen, R.V. Bowman, B.E. Clarke, E.E. Duhig, et al., Symbolic rule-based classification of lung cancer stages from free-text pathology reports, *J. Am. Med. Inform. Assoc*., **17** (2010), pp. 440–445.

[3] A. Nguyen, J. Moore, G. Zuccon, M. Lawley, S. Colquist, Classification of pathology reports for cancer registry notifications, *In: Health Informatics: Building a Healthcare Future Through Trusted Information-Selected Papers from the 20th Australian National Health Informatics Conference* (Hic 2012) **178**, (2012) 150.

[4] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K.S.J. Cooper, *et al.* Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model, *J. Biomed. Inform*., **42** (2009), pp. 937–949.

[5] M. Tanenblatt, A. Coden, I. Sominsky, The ConceptMapper approach to named entity recognition, *Language Resources and Evaluation, European Language Resources Association*, Malta (2010), pp. 546–551

[6] J. Sorace, D.R. Aberle, D. Elimam, S. Lawvere, O. Tawfik, W.D. Wallace, Integrating pathology and radiology disciplines: an emerging opportunity?, *BMC medicine*, *10*(1), (2012) 100.

[7] D. Martinez, L. Cavedon, Z. Alam, C. Bain, K. Verspoor, Text mining for lung cancer cases over large patient admission data, *Big Data Conference, Abstract Book. Big Data Conference, Melbourne April*. (2014), pp24-25.

[8] C. Bain, C. MacManus, Advancing data management and usage in a major Australian health service: The REASON discovery platform™, *Data Science & Engineering (ICDSE), 2014 International Conference on* (2014), 38–43.

[9] H. Han, W.Y. Wang, B.H. Mao, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *Advances in intelligent computing*. Springer Berlin Heidelberg, (2005). 878-887.

[10] A. R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *AMIA Annual Symposium Proceedings*, Washington DC, (2001) 17—21.

[11] W.W. Chapman, W. Bridewell, P. Hanbury, et al. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* **34** (2001).

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, Volume 11, Issue 1, (2009).

[13] J. Demšar, Statistical comparisons of classifiers over multiple data sets, *The Journal of Machine Learning Research* **7** (2006), 1–30.

[14] N. Thai-Nghe, Z. Gantner, L. Schmidt-Thieme, Cost-sensitive learning methods for imbalanced data, *in Proceeding of IEEE International Joint Conference on Neural Networks (IJCNN10)*, (2010).

[15] D.D. Lewis, Feature selection and feature extraction for text categorization. *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, (1992).