International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 6 | ISSN : 2456-3307

Text Mining Pathology and Radiology Records To Habitually Classify Against Disease : Computing The Control of Linking Data Sources

Dr. P. Radha[1], Mrs. B. Meena Preethi[2]

[1]Assistant Professor, PG & Research Department of Computer Science, Government Arts College, Coimbatore, India

[2]Assistant Professor, Department of BCA and M.Sc.SS, Sri Krishna Arts and Science College, Coimbatore, India

ABSTRACT

Text data mining, equivalent to text analytics is the course of action of deriving high-quality information from text. Text and data mining slot in recreation of obtaining insights from Health and Hospital Information Systems. Text mining system used for detecting admissions noticeable as positive for numerous diseases: Lung Cancer, Breast Cancer, Colon Cancer, Secondary Malignant Neoplasm of Respiratory and Digestive Organs, Multiple Myeloma and Malignant Plasma Cell Neoplasm, Pneumonia, and Pulmonary Embolism. Text mining explicitly inspects the effect of relating several data sources on text classification performance. Vector Machine classifiers for eight information resource combinations, and estimate using the metrics of Precision, Recall and F-Score. Sub-sampling techniques are used to address unbalanced datasets of medical records Radiology reports used as an initial data resource and add other sources, such as pathology reports and patient and hospital admission data, sequentially evaluate the research inquiry concerning the impact of the value of multiple data sources. Statistical significance is measured using the Wilcoxon signed-rank test. A subsequent set of experiments explores aspects of the system in greater depth, focusing on Lung Cancer. These tests tender improved understanding of how to optimum apply text classification in the context of imbalanced data of changeable completeness. Radiology questions plus patient and hospital admission data contribute valuable information for detecting most of the diseases, significantly improving performance when added to radiology reports alone or to the combination of radiology and pathology reports. The preference of the majority efficient combination of data sources depends on the precise disease to be classified. An approach whereby reports are electronically received and automatically processed, abstracted and analyzed has the potential to support expert clinical coders in their decision-making and assist with improving accuracy in data recording. Improving the cancer notifications process would provide significant benefits to oncology service providers, health administrators, clinicians and patients. The ultimate aim is to develop an automated system that can be trained to detect a new condition by having an expert in that condition analyse and annotate data directly.

Keywords : Health Informatics, Support Vector Machine, Handling Skewed Data

I. INTRODUCTION

Text mining is progressively more significant and authoritative 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 for enabling improved efficiency and communication within hospital systems, based on structuring of complex information for clinicians and hospital

administrators. The application of text mining and Natural Language Processing (NLP) techniques used to maintain important information in hospital proceedings are represented in free text format, e.g., radiology and pathology reports. Clinical records are a primary resource for disease-related information. However, the specific diagnosis associated with a visit may not be explicit in the records: the purpose of clinical investigations during a given patient visit or admission is often to help a clinician to decide on a disease diagnosis for a given set of signs and symptoms (in which case the disease is implied by the investigations, but not conclusively identified), or to select a treatment for a disease that has been previously identified (in which case the disease may not be directly mentioned in the visit record). A disease may be implicit in a given set of data, and only made explicit in free text fields in the record. Laboratory results in textual form, such as radiology and pathology reports, can provide strong clues as to the diagnosis, without stating a definite observed disease. Therefore, to classify clinical records related to a given disease, a promising strategy is to take advantage of these rich textual resources in a machine learning-based classifier. 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. However, the increase in data linkage (i.e., multiple data sources being linked by a unique patient identifier) in Hospital Information Systems is creating opportunities for more powerful and accurate text mining techniques that take advantage of information derived from multiple data sources [6]. In this paper, we examine the effect of linking multiple data sources on the performance of a text classification system for hospital reports. In particular, we describe performance in the context of the task of identifying patients admitted to a hospital for the treatment of a given disease, using various types of text-based records linked to patient admissions. The diseases we consider include Lung Cancer (a leading cause of death in developed countries) along with six others that are potentially confusable with that disease: specifically, we consider other cancers as well as non-cancer diseases centered in the chest area. The classifiers described in this paper are used to automatically map patient admissions to specific ICDs (International Code of Diseases) directly from hospital records. This task relates to automating ICD-coding – currently a time-consuming manual procedure at the core of the process followed to fund hospitals. Moreover, early and less onerous ICD-encoding can lead to improved disease surveillance and reporting, and support the rapid identification of patients who may be eligible for clinical trials. The focus of this paper is to evaluate the value of data linkage, and investigate the sources of value from among various possible hospital data sources. To this end, we consider a large collection of radiology and pathology reports, along with associated patient and hospital admission data (e.g., age, gender, length of stay), and build classifiers for each type of data source, as well as their combination. Several scenarios involving different diseases are tested, with separate binary classifiers built for each disease, and the statistical significance of the results is measured. The study demonstrates the value of heterogeneous medical data integration, contributes a deeper understanding of the contribution of specific data sources to the classification of hospital admissions, and proposes strategies for addressing the heavily skewed distribution of negative cases over positive ones. The effort required for information abstraction can therefore be an extremely labour and time intensive exercise.

II. RELATED WORK

Related works on reference resolution relevant to tumors or clinical findings have been the subject of several works. Coden et al identified co references in pathology reports using a rule-based system. Son et al

classified coreferrent tumor templates between documents with a MUC score of 0.72 precision and 0.63 recalls. Sevenster et al paired numerical finding measurements between documents. In Both radiology and pathology reports have been studied as a source of specific clinical information in previous text mining studies [1–7]. 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. Hripcsak et al.NLP techniques is usde evaluate the automatic coding of 889,921 chest radiology reports. Nguyen et al. [2] subsequent work used a rule-based system to classify cancer-modifiable pathology reports from a small corpus [3], obtaining very high sensitivity, specificity and Positive Predictive Value (PPV). Pathology\ reports have also been analyzed to study Skin Cancer [7], extract Breast Cancer characteristics into a knowledge model [4], and for training a named entity recognizer for relevant domain terms [5]. Bain and Mac Manus [8] survey in detail work on text mining cancer-related information from reports and records; we note that none of the works they list applies text mining techniques to multiple data sources simultaneously, as is done here. Linking (textual) medical records has previously been explored. For example, Heintzelman et al. [9] combined paper, electronic, radiologic, radiation therapy, and pathology medical records to analyse pain in prostate cancer patients. Zhao et al. [10] linked Electronic Health Records data to PubMed publications in pancreatic cancer prediction. However, to our knowledge, no prior work has systematically investigated the effect of linking multiple sources of clinical data on classification performance as we do in this work. In previous work , we presented a system for detecting Lung Cancer admissions based on radiology reports linked to radiology questions (which contain the purpose stated by the clinician for requesting a scan), pathology reports and patient metadata2 for a two-year period starting in July 2012. A similar approach is adopted in this paper, whereby we use text mining techniques to extract information about six additional diseases, and measure the statistical significance of classification performance using the different data sources. We also perform experiments with feature selection, and report on the impact of the amount of training data on learning. The specific goal of this paper is to explore the value of incorporating different data sources in classifying admissions against a selected set of ICD-10 codes. At least two independent groups have demonstrated that natural language processing can be as accurate as expert human coders for coding radiographic reports, as well as more accurate than simple text-based methods.

III. METHODOLOGY

3.1 SVMs and the skewed boundary

Support vector machines are based on the principle of Structural Risk Minimization from statistical learning theory. The idea of structural risk minimization is to find a hypothesis *h* for which we can guarantee the lowest true error. In the presence of noise, the idea of using a soft margin was introduced by Vapnik (1995). As noted earlier, data imbalance causes a default classifier to be learned which always predicts the "negative" class. Wu and Chang (2003) observed two potential causes for the problem of a skewed boundary:

1. The imbalanced training data ratio and
2. The imbalanced support-vector ratio.

For the first cause, we note that on the minority side of the boundary, the positive examples may not always reside as close to the "ideal boundary" as the negative examples. In terms of the second cause, consider the following: according to the KKT conditions, the values for $®i$ must satisfy $§ni=1®iyi=0$. Since the values for the minority class tend to be much larger than those for the majority class and the

number of positive support vectors substantially smaller, the nearest vicinity of an analysis peak is likely to be subjugated by negative support vectors. In other words, the decision function is more likely to classify a boundary point as negative. Table 1 lists the main record types (and number of records) in the platform that is relevant to our current task.

Table1. The numbers of records relevant to our study in REASON.

Data	Record numbers
Admissions	881,653
Pathology reports	667,303
Radiology reports	756,164
Radiology test orders	792,312

3.2 Analysis of Strategies for the Imbalanced problem for SVMs

To deal with the imbalanced boundary problem, several approaches were given for adjusting the skewed boundary.

- **Kernel transformation method**

Adaptively modifying the kernel function K based on the training data distribution is an effective method for improving SVMs. Amari and Wu (1999) propose a method of modifying a kernel function to improve the performance of a support vector machine classifier[11]. This method is based on the structure of the Riemannian geometry induced by the kernel function. The idea is to increase the separability between classes by enlarging the space around the separating boundary surface.

- **Biased penalties method**

Shawe-Taylor & Cristianini (1999) show that the distance of a test point from the boundary is related to its probability of misclassification. This observation has motivated a related technique which is used in his paper. The technique is to provide a more severe penalty if an error is made on a positive example than if it is made on a negative example. By using the cost factors and adjusting the cost of false positives and false negatives, such penalties can be directly incorporated into the SVM algorithm.

3.3 The REASON discovery platform for health informatics

The linked data for the studies described in this paper were extracted from the REASON Discovery Platform_ [8], a Health Informatics platform designed and implemented at Alfred Health (AH). AH is a major health service provider in metropolitan Melbourne, Victoria, Australia, which comprises three main health facilities, along with several smaller satellite facilities and many ambulatory services.

The REASON stage provides a particular information warehouse analysis of various data sources within the Alfred Combining heterogeneous sources of data is essential for accurate prediction of protein function[14].Health system, linked by unique anonymised patient id. Data was provided in the form of a de-identified set.

3.4 Data

For the purpose of this study, we extracted from the REASON platform the textual form of all radiology and pathology reports for a two-year period, from July 1, 2012 to June 30, 2014 (in accordance with Project 363/13 approved by the Alfred Health Human Research Ethics Committee). These reports were produced during normal treatment of patients within the hospital system.

Not all admissions are associated with all data sources, and some admissions may have multiple radiology or pathology reports (with a radiology question associated with each radiology report). Each report is assigned an admission identifier, which is in turn linked to patient metadata. The following metadata fields associated with each admission were extracted: patient's demographic data (e.g., gender, age, ethnic origin, country, language, etc.) 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). Time and date features were
abstracted to month and AM/PM value to alleviate data scarcity issues. The reason for admission field is a short free text field; it was presented as a bag of words. Each radiology report was also associated with a radiology question.
The initial number of reports and admission records used in this study was as follows:

- ✓ Radiology reports and radiology questions: 40,800;
- ✓ Pathology reports
- ✓ 20,872;
- ✓ Patient metadata: 121,700.

Each admission is associated with a set of ICD-10 codes, which are annotated in the admission record via a (human) clinical coder for reporting purposes. These codes are used in our study as ground truth labels to build the gold standard data sets for the selected diseases[13]. We note that the admission and coded data are expected to be of high quality, as this information is reported to the state Department of Health by the hospital for compliance and funding purposes. The data in the patient record is carefully reviewed and annotated by professional clinical coders in an internal manual process after the patient's discharge; it is hence suitable to treat these labels as having gold standard quality.

Each classifier is then trained to classify admissions against the ICD codes of interest, using contents of the associated reports and admission records. To evaluate the accuracy of the classifiers, the codes inferred by the classifiers are compared with the true ICD codes (which are removed from the data sets during experimental evaluation). The Victorian Department of Health mandates that the following prefixes are assigned to diagnosis codes to indicate condition onset [9]: P – Primary condition, C – Complication, A – Associated, and M – Morphology. These prefixes are assigned to each code in an admission. Fig. 1 illustrates a hospital admission record containing a radiology question, a radiology report, a pathology report, and metadata. The dashed square contains ICD-10 codes and their prefixes.

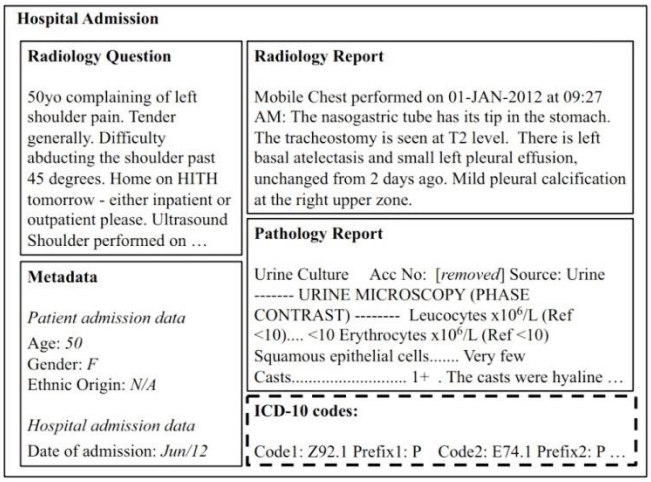


Figure 1. Hospital admission containing a radiology question, a radiology report, a pathology report, metadata, and a set of ICD-10 codes.

3.4 Techniques for handling skewed data

As noted earlier, the extracted data is heavily skewed, with relatively low numbers of positive occurrences of any given disease against total admissions . We performed subsampling to address the data imbalance problem due to the large number of negative cases for each disease. Specifically, we randomly selected a subset of negative admissions of approximately the same size as the set of positive examples. For example, when building the Lung Cancer classifier using only radiology reports (433 positive cases - Table 2), we randomly selected 433 negative samples to balance the dataset. To avoid bias to the specific selected subset, this process was repeated 10 times for each disease, using a different subset of negative admissions each time.[15]

Classification results reported below are averages of the results obtained across each of these 10 (balanced) subcollections, for each disease. We also analysed how models behave with an increasingly larger majority class: i.e., we investigated how classifier performance evolved across subcollections consisting of the full set of positive admissions for each disease and different (increasing) numbers of negative admission instances. The learning curve for these experiments is presented. In addition, we also investigated how linking the data influences classification performance by testing the model built using a subsample against all the remaining negative samples. Since all positive cases were used for training models, we report only Recall values for the negative class for this latest group of experiments. More complex alternative approaches to addressing the problem of highly skewed data are available this investigation is left for future work.

3.4.5. Evaluation

Evaluation of text mining and NLP systems typically employs the following three metrics: Precision, Recall and F-Score. Precision of positive (resp., negative) class is the ratio of correctly classified positive (resp., negative) values to the number of all instances classified as positive (resp., negative), this is also known as Positive (resp., Negative) Predictive Value. Recall of positive (resp., negative) class is computed as the number of correctly classified instances from the positive (resp., negative) class divided by the number of all instances from the positive (resp., negative) class; this is also known as sensitivity (resp., specificity). F-Score is the weighted harmonic mean of Precision and Recall. We performed 10-fold cross validation over each data subsample, whereby we randomly split data into 10 training/test sets. Instances in each fold were randomly selected; however, we ensured that the same random seed was used in each experiment, so that results were directly comparable. All steps of the algorithm that relied on the data (e.g., feature selection and training the SVM classifier) were performed inside the cross validation loop. (FIG2)

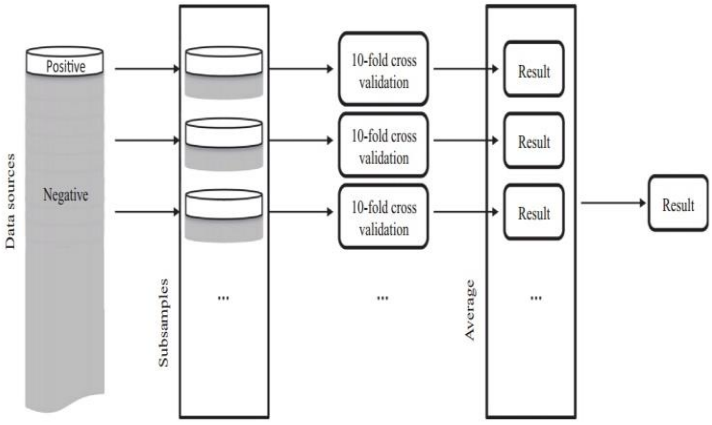


Figure 2. Experimental framework

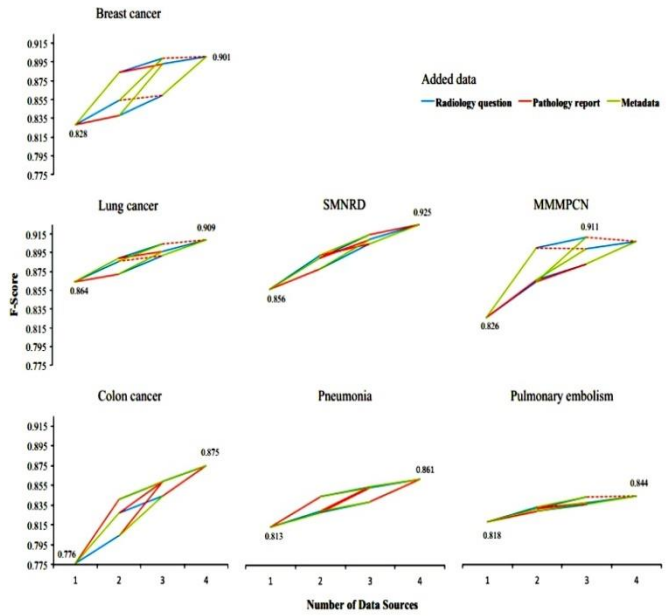
Table 2

Disease	ICD-10	# admissions with radiology reports	# admissions with pathology reports	# admissions wi metadata
Breast Cancer	C50	177	130	127
Colon Cancer	C18	186	154	163
Pulmonary Embolism	I26	368	264	351
SMNRD	C78	993	667	828
MMMPCN	C90	354	311	263
Lung Cancer	C34	433	340	396
Pneumonia	J12-J18	2,484	2,233	22,325

IV. RESULTS AND DISCUSSION

This section presents the results, grouped into the following two subsections: (a) main experiments over linked dataset and (b) additional experiments exploring performance under different configurations, performed on Lung Cancer only.

FIG3: data sources affect classifier performance



Our first set of results concerns the effects of incrementally adding new (linked) data sources on the performance of classifying admissions against the ICD-10 codes of the diseases of interest.

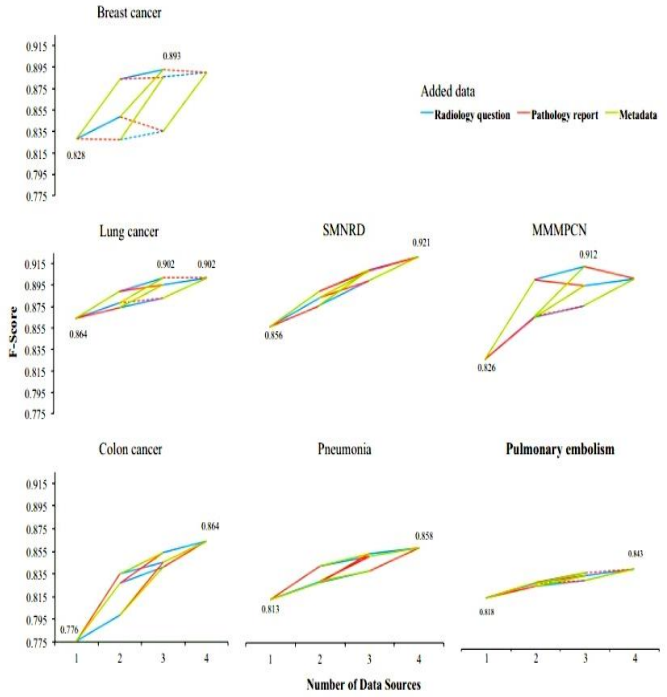


FIG 4: Results are shown with phrases distinguished by data source

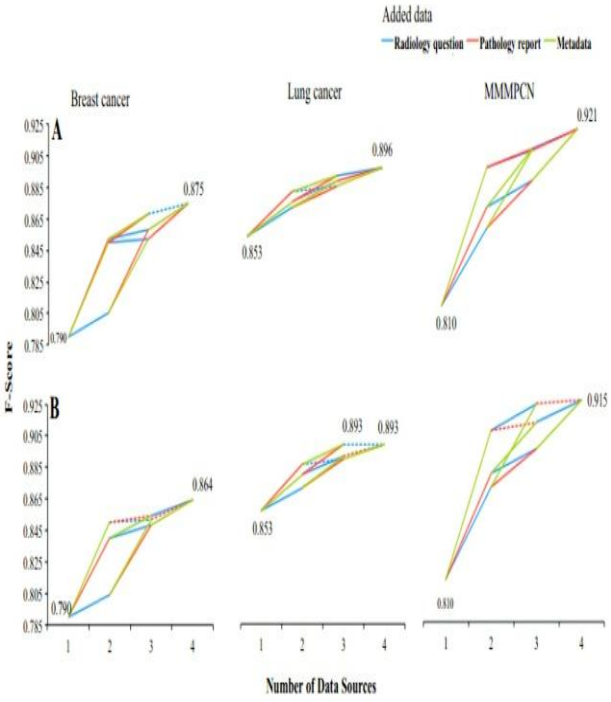


Figure 5. F-Scores for three selected diseases using only admissions containing both radiology and pathology reports

The charts in Figures 3 and 4 illustrate F-Score results for the seven diseases examined, using the two techniques for combining data sources. Figure 3 presents results without distinguishing phrases by data source (i.e., no prefixes on the features), while Figure 4 distinguishes the data source from which each feature was obtained (i.e., features are adorned with the data source). When following lines on single charts on both figures from left to right, we can observe how performance is affected by increasing the number of data sources. For example, the lower left corner on each chart shows performance using radiology reports only. The charts show that in most cases, as data sources are added, there is a statistically significant increase in performance. However, classifiers for three diseases (i.e., Lung Cancer, Breast Cancer and MMPCN) exhibited a decrease in performance specifically when adding pathology reports. This decrease was significant only for

MMPCN when using features adorned with the data source (Fig. 4). Analysis suggested that this was due to admissions without pathology reports being classified as negative due to an expectation of a pathology-specific feature: i.e., the classifiers did not distinguish between an admission containing no pathology report and one containing a pathology report with a lack of evidence for the disease. In other words, a negative admission with pathology reports could be treated the same way as a positive admission with no pathology reports, which would result in misclassification. We further explored the occasional performance degradations involving pathology reports by performing additional experiments using only those admissions containing both radiology and pathology\ reports. These additional experiments used the same techniques as the previous experiments. Fig. 5 shows the F-Scores obtained in the additional experiments by the classifiers for Lung Cancer, Breast Cancer and MMPCN. Complete results for all diseases. The top row in Fig. 5 shows results for the classifier built on merged data without the prefix tags (higher performance), while the bottom row shows results with the prefix tags attached to phrases to distinguish between data sources. Under both scenarios and for all diseases, optimal performance is obtained when all data sources are included

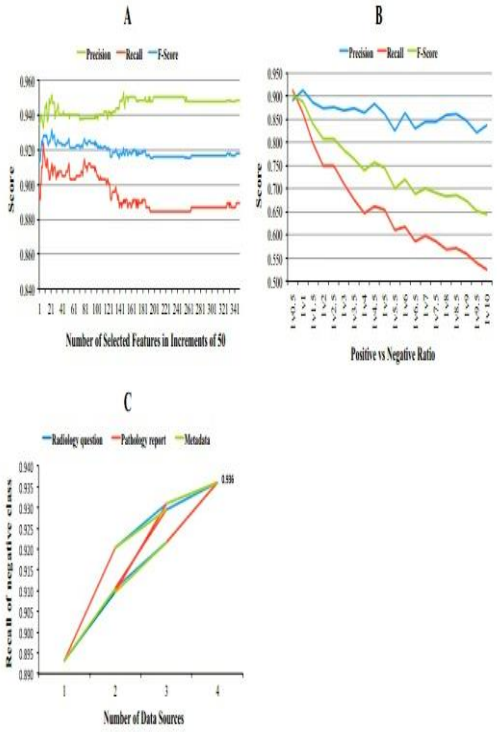


Figure 6. Additional experiments over the Lung Cancer dataset

V. CONCLUSION AND FUTURE SCOPE

Work has illustrated some of the challenges associated with building a successful automatic classification system for detecting admissions marked as positive for particular diseases. We have shown that mining multiple linked data sources improves classification performance on seven diseases considered. Radiology questions along with patient and hospital admission metadata seem to contain crucial information for accurately classifying most of the diseases, significantly improving performance when added to radiology reports or to the combination of radiology and pathology reports. In the future, we plan to explore alternative approaches to handling the data skewness. In addition, we plan to fine-tune the classifiers for specific diseases, and perform additional experiments with other ICD-10 codes. Main conclusions of the study may be presented in a short Conclusion Section.Author

should also briefly discuss the limitations of the research and Future Scope for improvement.

VI. REFERENCES

[1]. G. Hripcsak, J.H.M. Austin, P.O. Alderson, et al., Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports, Radiology 224 (2002) 157–163, http://dx.doi.org/ 10.1148/radiol.2241011118.

[2]. A.N. Nguyen, M.J. Lawley, D.P. Hansen, et al., Symbolic rule-based classification of lung cancer stages from free-text pathology reports, J. Am. Med. Inform. Assoc.17(2010)440 445, http://dx.doi.org/10.1136/jamia.2010.003707.

[3]. A. Nguyen, J. Moore, G. Zuccon, et al., Classification of pathology reports for cancer registry notifications, Stud. Health Technol. Inform. 178 (2012) 150– 156. <http://www.ncbi.nlm.nih.gov/pubmed/22797034 >.

[4]. A. Coden, G. Savova, I. Sominsky, et al., Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model, J. Biomed. Inform. 42 (2009) 937–949, http://dx.doi. org/10.1016/j.jbi.2008.12.005.

[5]. M. Tanenblatt, A. Coden, I. Sominsky, The ConceptMapper approach to named entity recognition, Proc. Seventh Conf. Int. Lang. Resour. Eval. Lr. (2010) 546–551.

[6]. J. Sorace, D.R. Aberle, D. Elimam, et al., Integrating pathology and radiology disciplines: an emerging opportunity?, BMC Med 10 (2012) 100, http://dx.doi.org/10.1186/1741-7015-10-100.

[7]. S. Kocbek, L. Cavedon, D. Martinez, et al., Evaluating classification power of linked admission data sources with text mining, in: R. Piskac, (Ed.), Annual Conference in Big Data in Health analytics (David Hansen 20 October 2015 to 21 October 2015). vol. 1468, 2015, pp. 1–7. <http://ceur-ws.org/Vol-1468/ bd2015_kocbek.pdf>.

[8]. C. Bain, C. Mac Manus, Advancing data management and usage in a major Australian health service, in: 2014 International Conference on Data Science & Engineering (ICDSE), IEEE, 2014, http://dx.doi.org/10.1109/ ICDSE.2014.6974609.

[9]. Online reference, Vic. Addit. to Aust. Coding Stand. Eff. 1July2015.<https://www2.health.vic.gov.au/Api/d ownloadmedia/{AFF11A9F-85A0-401C-ABE6-819167D9EC0A}> (accessed 1 Apr 2016).

[10]. I. Spasic , J. Livsey, J.A. Keane, et al., Text mining of cancer-related information: review of current status and future directions, Int. J. Med. Inform. 83 (2014) 605–623, http://dx.doi.org/10.1016/j.ijmedinf.2014.06.009.

[11]. C. Drummond, R.C. Holte, C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, in: Work Learn from Imbalanced Datasets II, 2003, pp. 1–8, doi: 10.1.1.68.6858.

[12]. B.X. Wang, N. Japkowicz, Boosting support vector machines for imbalanced data sets, Knowl. Inf. Syst. 25 (2009) 1–20, http://dx.doi.org/10.1007/s10115- 009-0198-y.

[13]. A.R. Aronson, Metamap: Mapping Text to the umls metathesaurus, Bethesda MD NLM NIH DHHS, 2006, pp. 1–26.<http://0-skr.nlm.nih.gov.library. law.suffolk.edu/papers/references/metamap06.pdf >.

[14]. A. Sokolov, C. Funk, K. Graim, et al., Combining heterogeneous data sources for accurate functional annotation of proteins, BMC Bioinformatics 14 (Suppl 3) (2013) S10, http://dx.doi.org/10.1186/1471-2105-14-S3-S10.

[15]. G.H.G. John, P. Langley, Estimating continuous distributions in Bayesianclassifiers, Proc. Elev. Conf. Uncertain Artif. Intell. Montr. Quebec, Canada 1(1995) 338–345. doi: 10.1007/b13634.