MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-1070

Automated Classification of Multi-Labeled Patient Safety Reports: A Shift from Quantity to Quality Measure

Chen Liang [a], Yang Gong [b]

[a] *Louisiana Tech University, Ruston, Louisiana, USA*
[b] *School of Biomedical Informatics, University of Texas Health Science Center, Houston, Texas, USA*

Abstract

Over the past two decades, there have seen an ever-increasing amount of patient safety reports yet the capacity of extracting useful information from the reports remains limited. Classification of patient safety reports is the first step of performing a downstream analysis. In practice, the manual review processes for classification are labor-intense. Studies have shown that the reports are often mislabeled or unclassifiable based on the pre-defined categories, which presents a notable data quality problem. In this study, we investigated the multi-labeled nature of patient safety reports. We argue that understanding multi-labeled nature of reports is a key to disclose the complex relations between many components during the courses and development of medical errors. Accordingly, we developed automated multi-label text classifiers to process patient safety reports. The experiments demonstrated feasibility and efficiency of a combination of multi-label algorithms in the benchmark comparison. Grounded on our experiments and results, we provided suggestions on how to implement automated classification of patient safety reports in the clinical settings.

Keywords:
Patient Safety, Machine Learning

Introduction

As emphasized in the Institute of Medicine's report 'To Err Is Human', error reporting and analysis are fundamental to patient safety [1]. Classification of patient safety reports is recognized as an initial step before any analysis and interventions can be applied [2]. However, the task of classification presents two urgent problems. Firstly, classified patient safety reports are of low quality due to sizable mislabeled reports and the reports under miscellaneous categories [3]. This problem may be caused by the lack of a deep contextual understanding of the reports. Secondly, conventional report classification is labor-intense and time-consuming. Presently, the classification task is largely completed by manual efforts. The rapid increase in the volume of the reports and research demands calls for an efficient and reliable solution for the classification task.

Patient safety reports are multi-labeled in nature. In many scenarios, an entity can be categorized with a single label. This is known as single-label classification. When an entity is associated with multiple labels, it is known as multi-label classification. Patient safety reports can be categorized with a number of pre-defined labels that fall into various categories such as incident types, type of harms, and contributing factors. The multi-labeled reports carry invaluable information for downstream analysis. Studies that evaluate patient safety reports tend to survey types and frequency of errors but draw less attention in the co-occurrence and relation of labels. These pieces of information may provide critical insights of reasoning about the root causes. In a study, researchers discovered that 42% of the reports were associated with more than two contributing factors. The increasing number of contributing factors per report may indicate the variation, increased complexity or severity of harm [4]. The multi-labeled patient safety reports can draw reviewers' special attention and thus can be better understood if they are well organized by categories.

Automated classification can be an alternative to the situation yet it presents interesting challenges when it is adapted to multi-label tasks. Automated classification is the task of using computers to learn associations between examples and labels. In-patient safety reports, an example can be an individual report in a corpus of reports. The classification is supervised because a classifier is built and trained by a set of prior labeled reports. Once the classifier is well trained, it can be used to predict candidate labels for unlabeled reports. As opposed to single-label classification, each report in multi-label classification may be associated with one or more labels. The multi-label classification is intuitive in human cognitive processes but creates extra complications in the computational process. Firstly, the training of a classifier is affected by many factors including the co-occurrence frequency of labels, hierarchical label relations, etc. [5; 6]. Multi-label algorithms must take these pieces of information into account but can easily lose feasibility or scalability by introducing sharply increased computational complexity [7]. Secondly, label imbalance is nearly inevitable in a multi-labeled corpus. As such, it negatively influences both classification performance and the selection of evaluation metrics. In specific, the prediction power of minority labels will decline since reports that are associated with these labels are less weighted during the training phase. Evaluation metrics such as exact match may not be as sufficient as it is in single-label tasks since a report may be predicted partially correct.

In general, multi-label algorithms can be categorized into two approaches: problem transformation, and algorithm adaptation. Problem transformation methods transform the multi-label problems to a number of single-label problems where a single-label problem can be solved by a range of single-label algorithms. A list of well-documented problem transformation algorithms includes binary relevance (BR), pairwise classification (PW), label combination (LC), and ranking and threshold (RT) [6]. Recently, a pruned sets method (PS) that is adapted from BR is reported to overcome many drawbacks of BR such as computational complexity [8]. A classifier chain method (CC) is also reported to improve LC for its less consideration of label relations, and sizable computational com-

plexity [9]. Algorithm adaptation methods modify specific algorithms to perform multi-label tasks, which are suited to a specific domain only.

In this paper, we adapted multi-label classification methods to categorize multi-labeled patient safety reports. The problem transformation algorithms were employed as the method to transform multi-label tasks to single-label tasks where binary classification algorithms serve as base classifiers. We present benchmark comparison between different combinations of algorithms. We argue that patient safety reports are multi-labeled and can be categorized by automated classification method with advanced efficiency.

Multi-labeled Reports

Patient safety reports contain crucial information to analyze incidents and to reduce recurrence. Reflected on a variety of factors mentioned in a report, an incident is multi-faceted. During the course of health care, a patient may have as many as hundreds of direct and indirect encounters with many persons, clinical procedures, and medical devices involved in hospital visits. Adverse outcomes may only occur in few encounters. But errors and contributing factors are likely to tie with many other encounters; especially those seemingly have no direct correlations to the patient harm. When an incident involves more than one error, the scenario is even more complicated. This situation is exactly the same as it is described in the Swiss cheese model of health care systems, which implies that medical errors are caused by simultaneous failures of multiple components in the system [10]. Fortunately, many of these factors can be extracted from the reports with either a clear mention or an indirect clue. Categorizing reports in a correct way has become a key to understanding both the root causes of a patient harm event and the risk factors of near miss or unsafe condition.

When it comes to massive reports, exploring multiple factors will provide more insights into the system's failure. Assuming that each factor represents a label, there are one or more labels associated with each report. The frequency of each label across the corpus can provide a general view of the distribution of labels. Another view of label distribution is the co-occurrence probability, which measures the probability of co-occurred labels. The co-occurrence probability is a piece of critical information in disclosing label relations. From the clinical perspective, if there is a high co-occurrence of a medication ordering system, a documentation error, and various drug reactions, for example, it is likely to have a malfunction or reoccurred user error relating to the medication ordering system. In a certain sense, therefore, label co-occurrence partially implies the potential factors that correlate with a given event.

Typically, the task of identifying multiple factors is a routine procedure of clinicians who review the reports. However, it requires a significant amount of time when the volume of reports scales up. The co-occurrence probability plays an important role in automated multi-label classification. It can be represented in the network visualization. In Figure 1, we generated a network graph based on a set of 2,919 patient safety reports. Each node in the graph represents a label that is used in the reports. The thickness of the edges that link a pair of nodes indicates the frequency of co-occurrence of two labels. For example, the label 'PerformanceFactor' has a high frequency in presence with either 'ProcedureTestTreatment' or 'MedicationIV'. The label 'Disability' hardly has any co-occurrence with other labels.

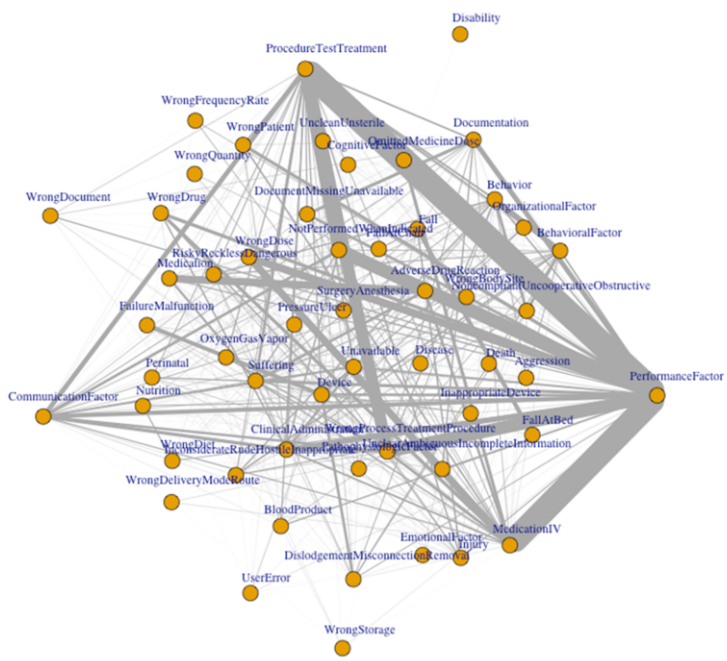

Figure 1 – Network visualization of co-occurrence probability of labels in 2,919 patient safety reports.

Automated Classification

In this section, we report our work to adapt multi-label text classification in the task of categorizing patient safety reports. The experiments were designed to evaluate the feasibility and efficiency of automated methods in multi-labeled reports.

Dataset

We used a corpus consisting of 2,919 de-identified patient safety reports from a university healthcare system. The original corpus contains 54 labels with a label cardinality of 2.89. The label cardinality is defined as

$$\boldsymbol{LCard(C)} = \frac{\sum_{i=1}^{N}|\boldsymbol{y_i}|}{\boldsymbol{N}} \quad (1)$$

N denotes the number of reports in the corpus C, y is the number of labels associated to with individual report, and L denotes the total number of labels. To avoid extreme imbalance of labels, we removed minority labels where each has less than 50 reports, resulting in 28 labels with a label cardinality of 2.58. See Table 1.

Table 1 – A demonstration of 28 labels in a hierarchy.

Top Level Category	Label	Frequency
Incident Type	Behavior	170
Incident Type	Clinical Administration	253
Incident Type	Device	317
Incident Type	(6 more)	-
Error Type	Adverse Drug Reaction	258
Error Type	Failure/Malfunction	105
Error Type	Fall at Bed	110
Error Type	(11 more)	-
Harm	Injury	123
Harm	Suffering	233
Contributing Factor	Behavioral Factor	169
Contributing Factor	Communication Factor	256
Contributing Factor	Performance Factor	1513

Procedure

We used problem transformation methods to solve the multi-label challenge. The problem transformation methods require a problem transformation algorithm in conjunction with a single-label algorithm that serves to build base classifiers. To perform the benchmark comparison, we chose a number of well-documented problem transformation algorithms considering both feasibility and computational complexity including BR, LC, RT, CC, and PS. Note that PW is not selected due to its significant computational complexity. For the parameter selection for PS, we used the optimized parameters (n = 0; p = {1, 3}) given that our corpus has a label cardinality of 2.58 and a number of 28 labels [8]. We also chose a number of single-label classifiers that represent a range of well-developed algorithms consisted of Naïve Bayes [11], Support Vector Machine (SVM) [12], k-Nearest Neighbor (kNN) [13], Decision Tree [14], and Decision Rule [15]. See Table 2.

Since these binary classifiers are not originally designed to process text data, we prepared our corpus as follows. (1) Snowball stemmer was used to reduce inflected terms to their root form [16]. (2) Rainbow list was used to remove stop words [17]. (3) Alphabetic tokenizer was used to break a string of text into terms. (4) Lower case token was applied to all the terms. (5) TF-IDF (term frequency-inverse document frequency) was used in the transformation of documents into a bag-of-words (BOW) matrix keeping 1000 unique terms [18].

Table 2 Single-label classification algorithms used in the experiments.

Algorithm	Implementation	Parameter
Naïve Bayes	NaiveBayes	
Support Vector Machine	LibSVM	Linear SVM
k-Nearest Neighbor	IBk	k = 1
Decision Rule	JRip	
Decision Tree	J48	

The experiments were performed on a 64-bit OS system with a processing power of 2.2 GHz with 4 Cores 8 Threads and a memory of 8 GB RAM. We used Python 3.0 to pre-process the corpus. The model training, evaluation, and statistics were performed on WEKA 3.6 [19] and MEKA 1.9.0 [20]. We employed 5 × 2 fold cross validation to randomize data and average results.

Evaluation Metrics

We consider both essential evaluation measures that are used in the single-label classification and the ones that are adapted for multi-label classification. The 0/1 Loss is a loss measure that assigns a '1' only if a label set is predicted exactly correct. The 0/1 Loss is defined as

$$0/1\ Loss = \frac{1}{N}\sum_{i=1}^{N} 1_{(\hat{Y}_i = Y_i)} \quad (2)$$

where $\hat{Y}$ denotes the predicted set of labels; Y denotes the exact set of labels.

Hamming Loss is the measure of labels that are incorrectly predicted. Instead of penalizing the incorrect match between two sets of labels, the Hamming Loss measures only the symmetrical difference between individual labels. Therefore, it is more forgiving than the 0/1 Loss. Hamming Loss can be referred as a partial match metrics whereas 0/1 Loss measures exact match. In multi-label classification, the form of Hamming Loss is defined as

$$Hamming\ Loss = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} 1_{(\hat{y}_{i,j} = y_{i,j})} \quad (3)$$

We also employed a multi-label accuracy measure that has been widely used in multi-label classification [21]. The form is defined as

$$Multi-label\ Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|} \quad (4)$$

The F measure is employed to provide a balanced metric between precision and recall. In multi-label classification, the generic F measure can take a number of forms depending on the different approaches of averaging metrics. In this study, we employed micro F measure. Mathematically, micro F measure favors labels with more documents, as it counts global metrics where labels with more documents have more weights. It is defined as

$$F_{Micro} = F\ Measure\ (y_{i,j}, \hat{y}_{i,j}) \quad (5)$$

In addition, we also consider the build time of models as a measure of computational complexity.

Results

We present the results by each evaluation metric. BR (M = 0.994) and LC (M = 0.994) performed slightly better than other multi-label algorithms but did not lead a clear difference in terms of 0/1 Loss. Naïve Bayes (M = 0.990) shows a slightly higher rate compared to other binary algorithms.

LC (M = 0.141) and $PS_{(p=3, n=0)}$ (M = 0.140) reveal an overall better performance in the measure of Hamming Loss. SVM (M = 0.136) shows the best performance among binary classifiers. The best combination appears to be the SVM in conjunction with RT (0.115).

BR (M = 0.110) is shown as the best problem transformation algorithm in the measure of multi-label accuracy. In the comparison of the base classifiers, Naïve Bayes (M = 0.110) outperformed the others. Naïve Bayes in conjunction with CC (0.128) appears to be the best combination.

BR (M = 0.195) reveals a higher F than other problem transformation algorithms. Naïve Bayes (M = 0.190) is the best base classifier overall. The best combination is Naïve Bayes in conjunction with CC (0.222). See Table 3 for details.

Table 3 – Micro F measure for different multi-label classifiers.

	NB	SVM	kNN	DR	DT	Rank
BR	0.212	0.184	0.152	0.219	0.206	1
LC	0.173	0.166	0.088	0.101	0.180	4
RT	0.185	0.118	0.122	0.190	0.192	3
CC	*0.222	0.197	0.092	0.209	0.165	2
PS $_{(p=1, n=0)}$	0.173	0.165	0.087	0.104	0.163	5
PS $_{(p=3, n=0)}$	0.173	0.163	0.084	0.111	0.157	6
Rank	1	3	5	4	2	

* best performance

$PS_{(p=3, n=0)}$ (M = 10.267 second) is the most efficient problem transformation algorithm as it shows the shortest build time. In terms of base classifiers, Naïve Bayes, SVM, and kNN (M = 0.228 second) cost lest build time. kNN in conjunction with $PS_{(p=1, n=0)}$ (0.015 second) is the most efficient combination.

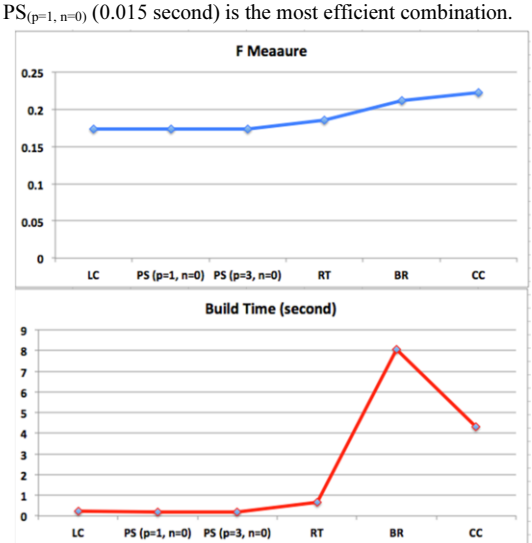

Figure 2 – F measure and build time for different problem transformation algorithms in conjunction with Naïve Bayesian.

We observed an interactive effect between F measure and build time over problem transformation algorithms. Figure 2 shows the comparison between different problem transformation algorithms when the base classifier is Naïve Bayes. CC appears to be a well-balanced algorithm considering its optimal performance and relatively short build time. Such an interactive effect did not show over binary algorithms.

Discussion

Multi-labeled Reports: from Quantity to Quality

This paper provided a novel perspective to investigate the nature of patient safety reports that is a report can be categorized by multiple labels. This argument corresponds to the Swiss cheese model that is frequently used as a metaphor of health care systems [10]. Errors are multi-faceted if they are viewed in a health care system where systems failures are preventable based on a number of relatively independent components. Such a component may refer to clinical administration, treatments, uses of medical equipment, etc., while each may produce unintended consequences. A systems failure occurs when by chance all the components produce errors. However, a systems view of these components can be obtained only if the reports are well categorized and, consequently, an aggregate analysis is made available.

The multi-labeled perspective of reports motivates a shift from quantity to quality measure. Patient safety studies have made remarkable progress, notably in constructing a nationwide reporting mechanism and ongoing focus to reducing harm by learning from lessons. These efforts have largely increased the number of reports, establishing a quantitative measure of patient safety such as the distribution of occurrences and frequency of errors. In fact, we are still far away from timely analysis and targeted quality improvement implied by the event reports. The vulnerability of health care system calls for special attention. We argue that one crucial gap is the limited understanding of reports, especially the intricate relations of the factors involved in a report. When the volume of reports increases, such pieces of relational information become more robust to indicate systems vulnerability. A prerequisite of performing quality measure as such is the capacity of extracting complex factors from massive reports. In this study, we suggested a multi-labeled approach.

Clinical Implementation

Our findings hold promise to improve the large-scale classification of patient safety reports. The experiments suggest feasibility and efficiency of using automated multi-label classification method to categorize patient safety reports. To balance the predictive power and efficiency, we found that CC in conjunction with Naïve Bayes is well performed. In addition, PS is also a promising method, as it largely reduces the model build time within a relatively small decline of predictive power.

More importantly, the multi-label classification is suited to the existing event reporting systems for improving the existing single-labeled classification and manual procedures. Clinicians who are responsible for case review and aggregate data analysis are expected to benefit from the automated classification. However, note that automated classification may not completely replace human effort for two reasons. Firstly, the automated classification results are not expected as good as manual results. The automated classification results should

serve as a reference source during the human review processes. Secondly, the automated classification has a limited capacity of predicting rare cases since it is generated based on the existing data. The rare cases, instead, may provide unique or crucial insights into medical errors.

With the scope of implementing automated classification in the reporting systems, we further suggest a uniformed classification scheme, which should provide a consistent and up-to-date classification hierarchy across health care providers. This is important to automated classification because the classification is supervised in a way relying on predefined labels to predict unlabeled reports. In the U.S., the Common Formats are widely used as a reporting guideline and classification schema in nationwide reporting. To extend the Common Formats' influence in guiding multi-label classification, additional work must be done to develop a classification hierarchy that supports automated classification directly, such as developing an ontological representation of patient safety reports [22].

Limitation and Future Direction

The label imbalance problem is a limitation, particularly in the context of multi-label classification. However, the label imbalance is inevitable in multi-label classification. The minority labels are even more common in medical corpus because medical entities (e.g., diseases, phenotypes, etc.) are not evenly distributed in a population. For example, minority labels can be 'death' or 'performance factor' in our corpus. To partially solve the problem, we removed extremely biased labels instead of creating synthesized reports. Consequently, it may lose some labels that are clinically important. Therefore, our approach still needs human guidance on categorizing minority labels at the current stage. In the future, we plan to enrich the corpus in both volume and sources.

We also note that the overall predictive power is comparatively small. Partially because we did not choose to perform feature selection and other manipulations that ought to boost the performance, as well as the fact that the 5×2 fold cross validation purports to find the most competitive classifiers under limited resources. The other interpretations may be the intricate semantic information and loss of information in the real-world patient safety reports. In the next step, we will investigate the effects of semantic information and domain knowledge in classification tasks.

Conclusion

The study demonstrated the effectiveness and efficiency of using automated multi-label classification on real-world patient safety reports. Our findings may improve (1) the process of understanding medical errors from an aggregate analysis, (2) clinical implementation of automated classification for large-scale patient safety reports.

Acknowledgements

The study is supported by a research grant (1R01HS022895) from the Agency of Healthcare Research and Quality, and the University of Texas System Grants Program (#156374).

References

[1] L.T. Kohn, J.M. Corrigan, and M.S. Donaldson, *To Err Is Human: Building a Safer Health System*, National Academies Press, 2000.

[2] L.L. Leape and S. Abookire, WHO draft guidelines for adverse event reporting and learning systems: from information to action, (2005).

[3] Y. Gong, Data consistency in a voluntary medical incident reporting system, *J Med Syst* **35** (2011), 609-615.

[4] P.J. Pronovost, D.A. Thompson, C.G. Holzmueller, L.H. Lubomski, T. Dorman, F. Dickman, M. Fahey, D.M. Steinwachs, L. Engineer, J.B. Sexton, and others, Toward learning from patient safety reporting systems, *Journal of critical care* **21** (2006), 305-315.

[5] Y. Labrou and T. Finin, Yahoo! as an ontology: using Yahoo! categories to describe documents, in: *Proceedings of the eighth international conference on Information and knowledge management*, 1999, pp. 180-187.

[6] G. Tsoumakas and I. Katakis, Multi-label classification: An overview, *Dept. of Informatics, Aristotle University of Thessaloniki, Greece* (2006).

[7] J. Read, Scalable multi-label classification, (2010).

[8] J. Read, B. Pfahringer, and G. Holmes, Multi-label classification using ensembles of pruned sets, in: *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 995-1000.

[9] J. Read, B. Pfahringer, G. Holmes, and E. Frank, Classifier chains for multi-label classification, *Machine learning* **85** (2011), 333-359.

[10] J. Reason, Human error: models and management, *BMJ* **320** (2000), 768-770.

[11] G.H. John and P. Langley, Estimating continuous distributions in Bayesian classifiers, in: *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995, pp. 338-345.

[12] C.C. Chang and C.J. Lin, LIBSVM: a library for support vector machines, in: *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM, 2011, p. 27.

[13] J. Wang and J.D. Zucker, Solving multiple-instance problem: A lazy learning approach, in: *Proc. 17th International Conf. on Machine Learning*, Morgan Kaufmann, 2000.

[14] J.R. Quinlan, Induction of decision trees, *Machine learning* **1** (1986), 81-106.

[15] W.W. Cohen, Fast effective rule induction, in: *Proceedings of the twelfth international conference on machine learning*, 1995, pp. 115-123.

[16] M.F. Porter, Snowball: A language for stemming algorithms, in, 2001.

[17] A. McCallum, Rainbow project, 1998.

[18] G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, *Information processing & management* **24** (1988), 513-523.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, The WEKA data mining software: an update, *ACM SIGKDD explorations newsletter* **11** (2009), 10-18.

[20] J. Read, P. Reutemann, B. Pfahringer, and G. Holmes, Meka: a multi-label/multi-target extension to weka, *Journal of Machine Learning Research* **17** (2016), 1-5.

[21] S. Godbole and S. Sarawagi, Discriminative methods for multi-labeled classification, in: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2004, pp. 22-30.

[22] C. Liang and Y. Gong, Knowledge Representation in Patient Safety Reporting: An Ontological Approach, *Journal of Data and Information Science* **1** (2016), 75-91.

Address for correspondence

Yang Gong, MD, PhD; Yang.Gong@uth.tmc.edu; 713 500 3547
7000 Fannin St. Suite 600, Houston, Texas, 77030, USA