Enhancing Patient Safety Event Reporting By K-nearest Neighbor Classifier

Chen Liang, MS, Yang Gong, MD, PhD
University of Texas Health Science Center at Houston, Houston, TX

Abstract

Data quality was placed as a major reason for the low utility of patient safety event reporting systems. A pressing need in improving data quality has advanced recent research focus in data entry associated with human factors. The debate on structured data entry or unstructured data entry reveals not only a trade-off problem among data accuracy, completeness, and timeliness, but also a technical gap on text mining. The present study suggested a text classification method, k-nearest neighbor (KNN), for predicting subject categories during event reporting. Our preliminary results demonstrated the feasibility of KNN classifier used for text classification and the advantage of such an application to raise data quality in reporting patient safety events.

Introduction

Much attention has been paid to how patient safety reporting system can improve the quality and safety of health service. Among many factors fundamental for a successful voluntary patient safety system, data quality has been a major concern. The incompleteness and inaccuracy of data were identified as two major problems of such systems. To solve these problems, researchers have found close associations between human factors and the system performance in terms of the completeness and accuracy of the data. Lacking of human factors was unfortunately found as a common shortcoming across patient safety reporting systems and has not been brought much attention over the past decade.

Data quality is greatly impacted by the process of data entry, which plays a critical role in healthcare information systems. The majority of patient safety data is recorded in free text. Although this might be an efficient and natural means for users to deliver an informative case, it could be costly to turn the raw information into a cognitively organized and manageable format for professionals to make utilization of it. Gong proposed the usage of pre-defined reporting categories as a key component in patient safety reporting system to improve data quality. However, structured data entry as such could be limited on timeliness and accuracy in both. Hua, Wang, and Gong introduced a new method using text prediction in unstructured data entry for balancing among accuracy, completeness, and timeliness in terms of data quality. The innovative use of this technique improved the system performance (13.0% time reduction and 3.9% accuracy increase), with the only concern being that the cognitive mechanisms underlying the prediction lists remain unveiled as the prediction list in the original study had to be manually prepared by domain experts. This may cause constraint on applying large-scale data entry in the real-world setting and therefore give rise to a new research topic.

One way to improve data entry is to incorporate both structured and unstructured data entry along with text classification. If the reporting system with this function can grow a list of categories from the imported data, which represents a taxonomy of the case, then we would be able to create a system that accepts unstructured data flexibly. Therefore, the system promises in generating quality data in terms of completeness. Our research question regards how to predict the event categories of a case presented in text document. Accordingly, we propose a text classification method used for predicting the taxonomy of event reports in free text. K-nearest neighbor (KNN) algorithm was employed to classify the event reports from Morbidity and Mortality Rounds on the Web (WebM&M). We chose KNN over other NLP techniques because the performance of KNN text classification can be improved to a large extent if the metric distance is under control and the training sets are well selected. KNN classifier allows us to categorize patient safety cases with minimized human effort. Unlike semi-structured the data extracted from electronic medical records (EMR), patient safety cases are fully narrative. For this reason, KNN is superior to other NLP techniques. The study aims at classifying free texted event reports into several categories using a well-trained KNN model. For example, a report can be manually labeled into a list of categories such as safety target, error type, clinical area, and so forth, and each category may contain several sub-categories. A set of
k-nearest training samples needs to be classified by domain experts and then used to train the model. Eventually, the model can proceed to classify unlabeled text documents automatically.

Method
We examined a total of 110 documents in the WebM&M where the patient safety cases are cross-labeled into 6 categories (safety target, error type, approach to improve safety, clinical area, target audience, and setting of care) with each category containing two or three levels of subcategories. The corpus we selected is under ‘safety target’, and is separated into two subcategories by domain experts, which ‘device-related complications’ (n = 30) and ‘diagnostic errors’ (n = 80). We used 70% of the corpus for training, and the rest for testing the model, which resulted in 18 ‘device-related complications’ documents and 59 ‘diagnostic errors’ documents as training sample, and 12 ‘device-related complications’ documents and 21 ‘diagnostic errors’ documents as testing sample. All the documents have been pre-categorized by domain experts when we extracted them from WebM&M, and these categories serve as the gold standard when testing the model predictions.

We implemented the KNN classifier in R along with ‘XML’, ‘tm’, ‘FNN’, and ‘plyr’ packages loaded. ‘XML’ was used for extracting free text data from the web. Other packages such as ‘tm’, ‘plyr’ and ‘FNN’ contain commonly used functions for data manipulation and KNN algorithms. The corpus was extracted from WebM&M and then passed a set of cleaning procedures as follows. (1) Punctuations were removed from the original text. (2) Words were toggled to lower case. (3) White space was stripped. (4) Stop words in English were removed from the text. When these procedures had finished, we then transformed the documents into a term-document matrix. All the documents were labeled with either ‘diagnostic errors’ or ‘device-related complications’ prior to being loaded in the model. This ensures a later comparison between predicted classification and the gold standard and consequentially allows the calculation on accuracy and F measure. The text classification was separated into two steps, which are training and testing. In the training phase, we gave a set of labeled documents so that the model can learn to map a list of correct categories to the corresponding documents. In the testing phase, the model tried to map a correct document to a certain list of categories. The training sample (70 %, 77 documents) was randomly selected from the pool (110 documents). By default, the number of the closest neighbors (k) is set to be 1. We tested the model performance by increasing the k from 5 at an interval of 5 until the F measure is out of capacity.

Results
The following figures (Figure 1) show the KNN classifier performance at a set of different k values. The top performance had k at 5 for both F measure (0.75) and accuracy (0.88), where the precision value is .86 and recall value is .67. We observed a decrease on performance as k increases from 5 to 20. F measure was no longer capable when k is larger than 20.

Figure 1. Performance of KNN classifier on different k. The left chart shows F measure. The right chart shows accuracy.
We observed a trade-off between a small number of neighbors (k) and large number of neighbors (k). When the KNN classifier has the best performance, a contingency table depicts its predictions per categories (Table 1).

Table 1. Contingency table per category.

<table>
<thead>
<tr>
<th></th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A1</td>
</tr>
<tr>
<td>Predicted</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>6</td>
</tr>
<tr>
<td>P2</td>
<td>3</td>
</tr>
</tbody>
</table>

A1 indicates ‘device-related complications’; A2 indicates ‘diagnostic errors’; P1 indicates ‘device-related complications’; P2 indicates ‘diagnostic errors’;

Discussion

As a simple-to-use method for classification, KNN used for text classification in patient safety cases has been evaluated effective 11-14. The reason we chose KNN algorithm is that it requires no training on the model itself, which renders it to be a simple algorithm to apply. In fact, KNN finds a neighbor by what language the neighbor speaks. This feature also better fits the situation in event reports, since some language will be used frequently when describing similar cases. In our study, the implementation of KNN classifier in safety event yielded a substantial improvement in automatic classifying the events. The model reached a fairly good performance where k was set to be around 5. With an accuracy of 88%, we are confident to use text classification technique for improving data quality in several ways. First, it can be used to classify text data with minimum supervision. As data grow rapidly, an outstanding reporting system requires the capability obtaining most useful information in a short time. As long as our model outputs with high accuracy, it can partially serve as domain experts’ role and therefore reduce the high cost of manual classification. Second, it allows the system to be flexible to accept unstructured format. Unstructured format has been used with a long tradition because it is easy for reporters to freely express. Nonetheless, reviewers of event reports also need an efficient and effective way to interpret and organize the reports in narrative format. Fortunately, our method holds the potential in solving the problem. Third, it sheds light on how to build an effective taxonomy of patient safety events and a sharable knowledgebase. Although the difficulty of reusing the event data in unstructured formats is challenging, text categorization provides a promising way to develop an ontological knowledgebase of patient safety events.

Data quality has been a barrier for patient safety event reporting system to play a pivotal role in improving healthcare quality. Issues with negative effects on data quality have been found associated with human factors engineering 15, 16. It has been a challenge balancing the completeness and efficiency of unstructured data entry and the effectiveness of structured data entry. To meet this challenge, it holds promise by adopting text classification as a data analytical technique. The major advantage of our method is thought to be emancipating the people, who use the data entry system to report cases, from arbitrary selections while still being able to extract useful information from the raw data. Natural language provides the richest information that conveys details of patient safety events 17, 18. On the other hand, it prevents traditional computerized system from effectively processing the data. It is generally agreed that manually categorize large-scale dataset is not practical. Similar to many other healthcare information systems, EMR has also been revealed conflict in benefiting from structured or free text data entry 19. We envision the application of text classification tends to extend to other systems in healthcare where human factors play a key role.

Although the KNN classifier revealed an impressive performance in the preliminary study, some limitations are worth to notice. The KNN algorithm is criticized due to its dependency on the selection of k value 11. That said, the best selection of k, which brings about the best performance on prediction, varies as the text documents differ. Therefore, text documents lacking completeness may not be a good use case. In addition, the sample size we tested in the present study is small and only contains two categories. In the
real-world situation, patient safety taxonomies usually have a hierarchical structure which contains three or four levels with each level having probably three to eight categories. Future study should evaluate different sorts of data with more categories.

**Conclusion**

The present study has demonstrated the feasibility of adopting KNN classifier to implement unsupervised text classification on patient safety reports. The positive results out of the model suggest the next step of research. In summary, this preliminary study shed light on how data quality could be improved by enhancing human factors in designing a healthcare information system supporting data analytics.

**Reference**