Filtering for Medical News Items Using a Machine Learning Approach

Wanhong Zheng M.B.B.S, M.C.Sc, Evangelos Milios* Ph.D, Carolyn Watters Ph.D.
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
Canada B3H 3W5

***Correspondent: eem@cs.dal.ca**

Abstract

In a recent paper we described work to filter medical news articles for targeted audiences. The approach was keyword based and one of the difficulties was extracting a feature set appropriate for the domain. This paper addresses the medical news-filtering problem using a machine learning approach. We describe the application of two supervised machine learning techniques, Decision Trees and Naïve Bayes, to automatically construct classifiers on the basis of a training set, in which news articles have been pre-classified by a medical expert and four other human readers. The goal is to classify the news articles into three groups: non-medical, medical intended for experts, and medical intended for other readers. While the general accuracy of the machine learning approach is around 78%, the accuracy on distinguishing non-medical articles from medical ones is shown to be 92%.

1. Introduction

As in many other fields, the introduction of the Internet and the World Wide Web throughout the 1990s has paved the way for significant advances in information exchange in the health area. Many health organizations including hospitals and government departments are now providing validated medical information to different communities. As an example, the American Medical Association published an online medical newspaper for American physicians[1]. As one of the main streams of web-based information, online news data are always of general interest and its volume is also tremendously large. In 1989, there were only 40 electronic newspapers and in 2000, the number had grown to more than 15,000[2]. In the meantime, like their paper format, many online newspapers contain a Health section to provide health related news of general interest. The result is that there is a clear filtering demand from myriad of news articles to retrieve the interested ones.

In general, electronic news articles on medical issues are helpful for everybody, from medical experts to laypersons without any medical training. Corresponding to individual background and interests, different readers prefer to select different news articles. In this paper we describe recent work on filtering medical news articles for targeted audiences.

The purpose of the research is to first identify the news articles with health or medical content and then categorize these articles by intended reader groups, layperson to medical expert. Using a machine learning approach, we address a news filtering service that can identify news items that are medical in nature and associate them with intended audience level. Two standard supervised machine learning techniques, Decision Trees and Naïve Bayes, were used to automatically construct classification rules on the basis of a training set, in which news articles have been pre-classified by a medical expert and four human readers with postgraduate university education but no medical training. The goal is to classify the news articles into three groups: non-medical, medical intended for experts, and medical intended for other readers.

2. Preliminary Work

Our news filtering approach is keyword-based. It identifies the keywords that characterize the medical content of the article. The first step is to filter out the non-medical news articles. The remaining articles were assigned MeSH (Medical Subject Headings) headings for content and classified into three readership classes: no particular medical background needed, medically knowledgeable and medical expert.

In order to attain the classification goal, a customized MeSH hierarchical vocabulary[3] was used as both a medical dictionary and classification taxonomy. A manual pre-defined weighting scheme on MeSH concepts was used to differentiate medical keywords of different readership levels. The extracted keywords were matched against the vocabulary and a MeSH sub-tree was created for each article. After such preprocessing, the next step is to apply supervised machine learning techniques that build classifiers based on a pre-classified training set.

3. Methodology

In this paper, we concentrate on two Machine Learning techniques: Decision Tree and Naive Bayesian classifier. Decision Tree generates clear descriptions of how the machine learning method arrives at a particular classification while the Naive Bayesian classifier was included for comparison purposes.

Decision Tree has been used to discover logical patterns within data sets for many years[4,5]. As a widely used approach to rule discovery[6], it can associate a class to an object based on tests applied to a set of attributes or features that describe that object[7].

A Naive Bayesian classifier[4] is a classifier based on Naïve Bayes Rule. In this supervised classification algorithm, Bayes Rule is used to induce the probabilistic connections between observed features and associated classes of news articles from previously observed training data. The observations are discrete feature vectors, with the assumption that the features used for deriving the prediction are statistically independent of each other and normally distributed for numeric attributes, although it is rarely valid in practical learning problems[8].

Despite its independence assumption, Naive Bayesian classifier performs well on many text classification tasks. It has been repeatedly shown to be competitive with more sophisticated induction algorithms[9]. For example, Clark and Niblett report Naive Bayes producing accuracy comparable to those for rule-induction methods in a medical domain[10]. For this reason, as a robust classifier, Naive Bayes Classifier can serve as a good comparison approach in our medical news classification system in terms of accuracy for evaluating other algorithms such as Decision Tree.

3.1. Training Data Preparation

All the training news articles were pre-assigned one of the three classes according to intended audience level: non-medical, layperson and medical expert. Totally 302 articles were used in the experiment. Since there were no gold standard criteria in classifying input news articles, the articles were first classified according to the sources they were from. Among those 302 articles, 100 articles were at expert level from The Doctor's Guide Website[11]; 102 articles for general readers were from the Health Section of Toronto Star[12] and Washington Post[13] on line newspapers; the remaining 100 articles non-medical ones were from Non-Health Sections of the same papers. To make the pre-classification more precise, in the second round the 102 articles in the second group were re-scanned by a medical expert and 45 articles were considered to be borderline and were extracted for further re-classification. Four human readers were invited to do the job.

As it is very difficult to avoid human bias, clear classification instructions were provided to the four human readers who did not have medical training.

The following is the criteria used for judgment:

The class is "medical expert level" if at least one of the following conditions is satisfied:

1) The reader cannot understand the article. Typical reasons might be: there are too many professional medical words in the article, the content is too hard to comprehend because the reader's medical knowledge is not enough.
2) What the article talks about is of no interest or potential interest to the reader.
3) The reader thinks this article is written for a medical doctor or a medical expert. That is to say, the ideal reader of this article should have fair amount of medical knowledge.

The class "medical for general readers" should be assigned if all of the following conditions are satisfied:

1) The reader can understand at least 90% of this article. And also he or she can explain it clearly in his or her own language to a regular layperson, say a middle school student or a construction worker who has no medical knowledge at all.
2) There should not be any "difficult" medical words that the reader needs to look up in the dictionary or the Web.
3) The content of the article attracts the reader's interest.

Of the 45 articles, 14 articles were classified with 2/2 votes and therefore were discarded as having an ambiguous classification. 17 articles were re-classified to a different class and the remaining 14 were classified back to the same group.

Table 1 shows the summary of the training data pre-classification result.

Pre-Assigned Class	Source	Number
Non-medical	Online newspaper (Non-Health Section)	100
Medical for general readers	Online newspaper (Health Section)	71
Medical expert level	1.www.docguide.com 2.Online newspaper (Health Section)	117

Table 1. User-labeled Training Data Summary

3.2. Optimal Feature Selection

The pruned MeSH vocabulary was subjectively weighted by the same medical expert. Each term or

MeSH concept was assigned a weight indicating its degree of medical relevance for classifying news articles: Level 1 term for medical expert; one example of this level is “Douglas Pouch”; Level 2 for medically knowledgeable people, for example, “Pelvis”; Level 3 medical term for laypersons such as “Stomach”, and Level 4 non-medical term like “back”. The reason we have Level 4 in MeSH is that some MeSH concepts like “back” and “neck” can be either medical or non-medical depending on the context.

The first feature set used in the experiment consists of six features. The first four features are the fraction of Level 1, Level 2, Level 3, Level 4 words in the article respectively. The fifth and sixth features are the fraction of Level 1, Level 2 and Level 3 words combined in the text of the article and in the title of the article respectively.

The second feature set is a modification of the first group. It consists of seven features. The first three features are still the fraction of Level 1, Level 2, and Level 3 in the article. The fourth and fifth features are the fraction of Level 1 words in the total medical words (Level 1, Level 2 and Level 3 words combined) and Level 1 and Level 2 words combined in the total medical words. The last two features are same as the fifth and sixth features of the first group.

In summary, all the attributes used are listed as follows and Table 2 shows the selection of the attributes for each group:

I. Level 1 words/Total words in the article
II. Level 2 words/Total words in the article
III. Level 3 words/Total words in the article
IV. Level 4 words/Total words in the article
V. Level 1 words/Total medical words (Level 1,2,3 words)
VI. Level 1,2 words/Total medical words in the article
VII. Level 1,2,3 words/Total words in the article
VIII. Level 1,2,3 words in title/Total words in the title

Group Number	1	2
Attributes Used	I ~ IV, VII, VIII	I ~ III, V ~ VIII

Table 2. Attribute Selection for Experiments

4. Experimental Results

The estimate of classification accuracy is obtained using stratified ten-fold cross-validation. In this procedure, the training examples are randomly divided into ten equal-sized partitions. Each partition, which preserves the original class distribution, is used in turn as test data for the decision tree trained on the remaining nine partitions. This approach entails that less data is available for building the model, but the quality of the estimate of the accuracy is improved, because it is based on unseen data, i.e. data not used for training. In this case, as is shown in Table 3, the correctness is around 77% for both Decision Tree using 6 attributes and the Tree using 7 attributes.

	Decision Tree 1 using 6 attributes	**Decision Tree 2 using 7 attributes**
Correctly Classified Instances	77.08%	77.43%

Table 3. Decision Tree Performance Comparison

Note that the classification accuracy for stratified cross-validation shows no difference between two attributes selections. If we further investigate the confusion matrix of the two decision trees, we can find that our system works well on differentiating medical articles from non-medical ones. This can be demonstrated by the data from Table 4 that shows the stratified cross-validation confusion matrix of Decision Tree 1 using 6 attributes and Table 5 that shows that of Decision Tree 2 using 7 attributes.

Note here that both classifiers perform well on non-medical articles. The accuracy is consistently 91% for both Decision Tree 1 using 6 attributes and Tree 2 using 7 attributes.

Further investigation on confusion matrices shows that the classifier also achieves high accuracy for medical-for-expert group. The accuracy is 79% (93/117) and 86% (101/117). But for medical-for-layperson articles, the Decision Trees have correctness of 55% (39/71) and 46% (33/71).

DT-assigned **Pre-assigned**	**Medical expert**	**Layperson**	**Non-medical**
Expert	**93**	23	1
Layperson	25	**39**	7
Non-medical	2	8	**90**

Table 4. Stratified Cross-validation Confusion Matrix for Decision Tree 1 on User-labeled Training Data Using 6 Attributes

DT-assigned **Pre-assigned**	**Medical expert**	**Layperson**	**Non-medical**
Expert	**101**	16	0
Layperson	32	**33**	6
Non-medical	2	9	**89**

Table 5. Stratified Cross-validation Confusion Matrix for Decision Tree 2 on User-labeled Training Data Using 7 Attributes

The second Machine Learning technique being used is Naïve Beyes (NB) Classifier. The data from Naïve Bayes classifiers is consistent with that from the Decision Tree. Table 6 shows the correctness for two Naïve Bayes classifiers using six and seven attributes respectively. Both classifiers can classify 77 - 79% articles correctly.

Naïve Bayes Classifier Data	NB 1 (with 6 attributes)	NB 2 (with 7 attributes)
Total Instances	288	288
Correctly Classified Instances	77.43%	78.47%

Table 6. Naïve Bayes Classifiers Data Summary

The confusion matrices of the two Naïve Bayes Classifiers (Data not shown) further proved that the system performs well on non-medical articles. Both Naïve Bayes classifier using six attributes and Naïve Bayes classifier using seven attributes can classify 93% articles correctly. Moreover, the data for the other two classes is also consistent with that from two decision trees. For Medical for Laypersons, Naïve Bayes classifier using six attributes has correctness of 54% and Naïve Bayes classifier using seven attributes 57%, and for Medical expert group, the correctness is 88% and 87.5%. The data is consistent with the accuracy of two Decision Trees on these two groups of articles. Note that in Naïve Bayes classifiers, using six attributes doesn't perform better than using seven attributes on medical-for-expert articles, which is consistent as in Decision Tree classifiers.

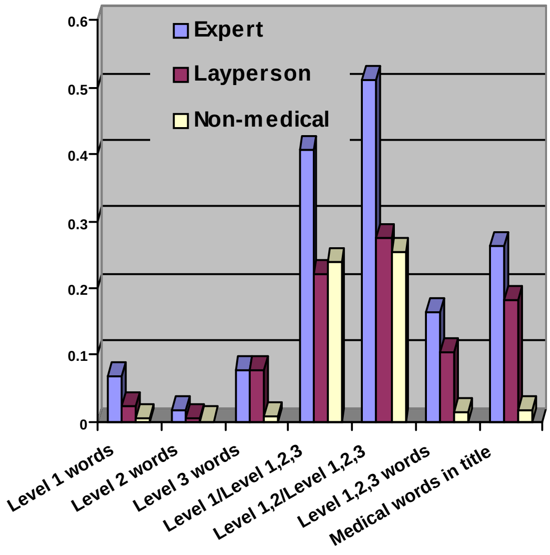

Figure 1. Attribute Mean Comparison in Naïve Bayes Classifier with 7 Attributes

Figure 1 shows the comparison of the mean values of the seven attributes used in Naïve Bayes classifier 2. Except the attributes "Level 2 words" and "Level 3 words", all other five attributes show some differences between "medical-for-expert" and "medical-for-layperson" articles. This means that these five attributes can be used for final class determination, which is consistent with the Decision Tree 2 that uses seven attributes: the tree uses all five attributes as features after pruning (Data not shown). On the other hand, the mean distribution of the seven attributes in three classes shows that five of the seven chosen attributes show great difference between medical and non-medical articles. But the difference is small for two others – Level 1 words/Level 1,2,3 words and Level 1,2 words/Level 1,2,3 words.

We tested the Decision Tree classifiers with test articles pre-classified by a medical expert. Table 7 shows the correctly classified instances in each class for two Decision Tree classifiers.

Class Classifier	Non-medical	Layperson	Expert
Decision Tree with 6 attributes	9/10	8/10	10/10
Decision Tree with 7 attributes	9/10	8/10	9/10

Table 7. Decision Tree Testing Result Summary

Both Decision Trees predict more than 90% of non-medical and medical-for-expert articles correctly. The correctly classified medical-for-layperson articles are also 80%. Although the number of test articles is not sufficient enough to show the system performance, this is a further indication that the classification system is good for distinguishing medical articles from non-medical ones.

5. Conclusion

The general effectiveness of the Decision Tree learners on stratified cross-validation from user-labeled training data is around 77% (Table 3), which is confirmed by the Naïve Bayes classifiers. However, the data in Table 4 and 5 show that the performance on non-medical articles is close to 92%. Therefore although the general accuracy is not high, machine learning is a good approach to do binary classification for medical and non-medical identification. And the performance shows no remarkable difference between the two attributes selections.

6. Acknowledgement

The work of this paper has been supported in part by the Natural Sciences and Engineering Research Council of Canada. Many thanks to Dalhousie University WIFL (Web Information Filtering Lab) members and especially those who have spent their

precious time reading and labeling numerous news articles that are used as training data.

References

[1] American Medical News. Online at: [http://www.ama-assn.org/public/journals/amnews/amnews.htm] Last Accessed: 2002/1/12

[2] Abyz Web Links. 2001. Online at: [http://www.abyznewslinks.com] Last Accessed: 2002/1/12

[3] Medical Subject Headings. Online at: [http://www.nlm.nih.gov/mesh/meshhome.html] Last Accessed: 2002/2/16

[4] Weka 3 – Data Mining With Open Source Machine Learning Software in Java. 2001 Online at: [http://www.cs.waikato.ac.nz/ml/weka/] Last Accessed: 2001/10/12

[5] J. R. Quinlan. "C4.5: Programs for machine learning". Morgan Kaufmann, San Mateo, California, 1993.

[6] Information Discovery, Inc. Rules Are Much More Than Decision Trees. Online At: [http://www.datamining.com/trees.htm] Last Accessed: 2002/1/10

[7] S. R. Safavin & D. Landgrebe. "A survey of decision tree classifier methodology". IEEE Transactions on Systems, Man and Cybernetics, 21(3): 660--674, May/June 1991

[8] Frank E., Trigg L., Holmes G. and Witten I.H. (2000). "Technical Note: Naive Bayes for regression". Machine Learning, 41(1) 5-26, October.

[9] G. H. John and P. Langley (1995). "Estimating Continuous Distributions in Bayesian Classifiers". Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. 338-345. Morgan Kaufmann, San Mateo.

[10] P. Clark, & T. Niblett (1989) "The CN2 induction algorithm", Machine Learning 3(4), 261-83.

[11] Doctor's Guide. 2002. Doctor's Guide Home Page. Online at: [http://www.docguide.com/] Last Accessed: 2002/1/12

[12] Toronto Star. 2002. Online at: [http://www.torontostar.com/] Last Accessed: 2002/1/12

[13] Washington Post. 2002. Online at: [http://www.washingtonpost.com/] Last Accessed: 2002/1/12