Classifying questions on an online cancer forum

Weiling Li
324 IST building, University Park, Penn State University
State college, PA 16802
wul135@psu.edu

Abstract
Online forums are an important source of social support for cancer survivors and their informal caregivers. To organize large collections of medical posts for efficient retrieval is thus very important. In this project, we automatically classify questions based on a proposed hierarchical architecture. Classification based on features selected from high term frequency words is the baseline. And more features are added. A large, robust cancer terms resource, namely, National Cancer Institute (NCI)², is used to help to learn clinical features, and subjectivity lexicons (strong or weak)³ provided by University of Pittsburgh to classify objective questions, asking for emotional support (strong subjectivity) and opinion (weak subjectivity). Several popular classifiers are chosen in order to get the better accuracy.

1 Introduction
Cancer accounts for 1.8 million deaths in China, 0.57 million in the U.S. (both in 2010), and 7.6 million deaths worldwide in 2008 [1]. As of 2007, 11.7 million Americans have been diagnosed with cancer and are either living free of cancer or still have evidence of the disease [1]. Social support, in both clinical and emotional ways can help cancer survivors cope better with the disease and improve their life. Moreover, peer support and information can be quite valuable, since cancer survivors’ experiences are unique and may not be well understood by their family members, friends or care providers.

As large collections of questions exist in the forum, we need the question classification to assign a given question to predefined question categories in information retrieval or question-answering system on a better performance. In the medical domain, physicians have classified questions based on clinical knowledge [2]. As my knowledge, cancer questions classification has not been done before. To validate our method, we use the data from the online forum in the American Cancer Society Cancer Survivors Network® (CSN). CSN (http://csn.cancer.org) is a community of more than 146,000 registered members created by and for cancer survivors and caregivers. Its online forum consists of 38 discussion boards.

1 Course: CSE 583. Advisor: Dr. Liu. Customer: Dr. Mitra
2 http://www.cancer.gov/dictionary
3 http://www.cs.pitt.edu/mpqa/subj_lexicon.html
2 Question Types

Many clinical question classifiers are based on Ely’s evidence taxonomy to category answerable and unanswerable questions [2]. Evidence bases questions are good questions that can be answered by evidence from scientific and medical research for better caring patients; on the other hand, a patient-specific question is usually too vague to offer an answer [3]. Moreover, evidence-based question can be further classified for generating question-type-specific answer. No intervention questions are factoid questions that usually answered by short answers, on the other hand, intervention questions are scenario-based questions that usually answered by long and complete answers [3].

Online social support is different from normal question-answering system, always involves emotions [4]. Objective and subjective questions are two important styles of questions in the medical forum. Objective questions are expected to be answered by shorter answers from experts. Emotional support questions are involved stronger objectivity and expect positive responses, while opinion queries expect less emotion involved, and answers can be either positive or negative or both or neutral.

A new question hierarchy is proposed in Figure 1. The hierarchy is a 3-level binary tree, where in the 1st level, questions are divided by clinical questions and companionship questions, and the 2nd level divides clinical questions into question asking for factual information or subjective questions, and subjective questions are further divided into questions asking for opinion/advice and emotional support in the 3rd level. In other words, questions in the online cancer forum can be divided into 4 types: factual information, opinion/advice, emotional support, and companionship. The followings are definitions and examples of each type.

![Figure 1. Questions types’ hierarchy](image)

Facts (include referrals, direct questions, etc.)

• Requested – Inquiring about a source for information or assistance. Examples include the following:
  – Do you know of a site I might find some answers to my particular case.
  – Do any of you know an endocrinologist or onco that knows or understands MTC (medullary thyroid cancer)?
  – Where do you buy digestive enzymes and what are they called?

Opinion / Advice

• Requested – Asking for opinions on some topic. Examples include the following:
  – How is hospital X for Y cancer?

• Requested – Asking for guidance or advice. Examples include the following:
She will appreciate any advice.
Any/all suggestions welcome.
I was wondering if there is any other way to prevent this.

*Response can be positive / negative*

**Emotional Support** (includes understanding/empathy, and encouragement)
*Requested – Searching for someone who has an appreciation for or can identify with one’s experience or feelings often. Examples include the following:
I wonder if others who have had the same surgery as I have are experiencing pain? Please let me know.
•Requested – Expressing despair and/or the need to feel hopeful or courageous. Examples include the following:
It makes me very sad to think that Mom will not be here to see my son grow up. How can I deal with it?
•Always expect for positive response

**Companionship**
It consists of messages that bestow a sense of being connected to a group. These are unrelated to cancer. Following are the examples
•Chatting: “Anyone joining me for the concert”?
•Messages containing jokes, poking fun at another member, humour etc

3 Data Description

The whole raw dataset includes posts from July 2000 to October 2010, comprised of 48,779 threaded discussions with more than 468,000 posts over 38 discussion boards (25 cancer-specific, and 13 others) from 27,173 users. 217 questions are randomly selected from start posts as the dataset, including 55 factual questions, 51 opinion/advice questions, 99 emotional support questions, and 12 companionship questions.

RapidMinder\(^d\) is used to do text processing. After doing tokenization, filter stop words, porter stemming, 945 words (features) are left in the dataset. Using PCA (princomp in Matlab), visualization of the 1\(^{st}\) and 2\(^{nd}\) principal components is in Figure 2.

![Figure 2. PCA visualization of the dataset](image)
4 Baseline

Term frequency (TF) is used to filter features. After 10 times 10-folder cross validation via bagging (Treebagger with ntree=100 in Matlab), testing accuracy is shown in Table 1.

<table>
<thead>
<tr>
<th>TF</th>
<th>≥2</th>
<th>≥3</th>
<th>≥4</th>
<th>≥5</th>
<th>≥6</th>
<th>≥7</th>
<th>≥8</th>
<th>≥9</th>
<th>≥10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accur acy</td>
<td>0.5912 (±0.073)</td>
<td>0.6025 (±0.077)</td>
<td>0.6176 (±0.107)</td>
<td>0.6178 (±0.087)</td>
<td>0.6298 (±0.097)</td>
<td>0.5919 (±0.108)</td>
<td>0.5781 (±0.109)</td>
<td>0.5869 (±0.098)</td>
<td>0.5430 (±0.098)</td>
</tr>
</tbody>
</table>

Using the function OOBPermuttedVarDeltaError provided by Matlab, feature importance can be gotten. With the condition of TF≥6 and OOBPermuttedVarDeltaError>0.2, 16 features are always existed when running 10 times. Mean accuracy is further improved to 0.6447 with standard deviation = 0.0779, and mean accuracy for each class is: accuracy1 = 0.8430, accuracy2 = 0.5530, accuracy3 = 0.654, accuracy4 = 0 (because of imbalanced data).

5 More Features

Dictionary of cancer terms in the National Cancer Institute (NCI) helps to learn clinical terms (features), and 153 clinical words/phrases (features) are learned. 127 subjectivity words (43 words with weak sentiment that are related with opinion/advice questions, 84 words with strong sentiment that are related with emotional support questions) are learned according to MPQA. After observing features and questions, two more features are added (“recommend” related with opinion questions, and “experience” related with emotional support questions).

Using the total of 298 features, bagging is redone. With the condition of TF≥5 and OOBPermuttedVarDeltaError>0.2, the optimal accuracy is gotten. 74 features are selected, and mean accuracy is 0.6855 with standard deviation = 0.0862. And the accuracy of each class is: accuracy1 = 0.8200, accuracy2 = 0.5380, accuracy3 = 0.7596, accuracy4 = 0.0700. In other words, the addition of features improves the accuracy rate a little bit.

6 Feature Selection and Classification

6.1 Feature Ranking

Augmented variance ratio (AVR) is used for feature ranking. In [5], for a feature F with values $S_F$ in a data set with C total classes, the augmented variance ratio (AVR) is calculated as

$$AVR(F) = \frac{\text{Var}(S_F)}{\left(\frac{1}{C} \sum_{i=1,...,C} \left(\text{mean}_i(S_F) - \text{mean}_0(S_F)\right)^2\right)}$$

where meani(SF) is the mean of feature F’s values in class i and Vark(SF) is the variance of the subset of values from feature F which belongs to class C. This ratio is the ratio of the variance of the feature within subjects, with an added penalty for features which may have small intra-class variance but which have close inter-subject mean values.

As we have just two classes, we can reformulate the augmented variance ratio function as follows for easier computation.

$$AVR(F) = \frac{2|\text{mean}_0(S_F) - \text{mean}_1(S_F)| \cdot \text{var}(S_F)}{\text{var}_0(S_F) + \text{var}_1(S_F)}$$

http://rapid-i.com/content/view/281/225/

4 advice, cry, depress, doctor, else, feel, get, heard, hospital, life, mean, old, opinion, suggest, time, tumor
6.2 Feature selection

Two key components of feature selection are:
1) a search strategy to go through the feature list and select a subset of it and
2) an evaluation function to measure the discriminative power of the selected features.

The main idea is to select a subset of the attributes that are influential in the classification task.

Sequential Forward Selection (SFS)

Sequential Forward Selection is the simplest greedy search algorithm, starting with the empty set, sequentially adding the feature \( x^+ \) that results in the highest objective function \( J(Y_k+x^+) \) when combined with the selected features \( Y_k \). Update \( Y_k + 1 = Y_k + x^+ \) and \( k = k + 1 \). When no additional feature improves the objective evaluation function or when no more feature left to be selected or when a pre-set size of feature subset is reached, the SFS stops.

The algorithm is written as follows:
1. let \( P = \emptyset \) be the current set of selected features
2. let \( Q \) be the full set of features
3. while size of \( P \) < a given constant
   (a) for each \( v \in Q \)
      i. set \( P' = \{v\} \cup P \)
      ii. train the model with \( P' \) and keep the validation performance
   (b) set \( P = \{v^*\} \cup P \) where \( v^* \) corresponds to the best validation performance obtained in
      step 3a
   (c) set \( Q = Q \setminus \{v^*\} \)
   (d) keep the validation performance obtained with current \( P \)
4. return the best set \( P \)

SFS is good when the best feature subset is relatively small; while its advantage is that once a feature is selected, it cannot be removed again.

When data is imbalanced, maximizing (sensitivity + specialty) is one of the criterions, the reason to choose the sum of sensitivity and specialty to indicate performance. When data is relatively balanced, \textit{avg. accuracy} is one of the criterions.

6.3 SVM

Support vector machines (SVMs) have considerable potential for supervised classification analyses, but their binary nature has been a constraint on their use in remote sensing. This typically requires a multiclass analysis be broken down into a series of binary classifications, following either the one-against-one (1A1) or one-against-all (1AA) strategies. In this project, classification follows a binary tree, 1A1 SVM is chosen to do classification in each level. Default SVM (with linear kernel function) in Matlab is used.

7 Experiments

10 times 10-fold cross-validation with the support vector machine (SVM) classification is used for supervised learning. Accuracy, sensitivity, and specificity are tested to analyze performance. Based on the last subsection, top features ranked by AVR are as input to the SFS, and average accuracy/ or maximizing (sensitivity+speciality) is one of the criterions.

In the 1st level, all 298 features are used to classify clinical and companionship questions. When maximizing (sensitivity+speciality),

- mean(accuracy)= 0.6977 std(accuracy)=0.1237
- mean(error1)= 0.3171 std(error1)= 0.1385
- mean(error2)= 0.000000 std(error2)= 0.000000
- mean(sensitivity)= 0.6829 mean(specificity)= 1.00000
However, clinical questions are more important to us. When maximizing accuracy, on average, 30 features are selected.

\[
\text{mean(accuracy)}=0.954113 \quad \text{std(accuracy)}=0.030317 \\
\text{mean(error1)}=0.000000 \quad \text{std(error1)}=0.000000 \\
\text{mean(error2)}=0.8332 \quad \text{std(error2)}= 0.351638 \\
\text{mean(sensitivity)}=1.000000 \quad \text{mean(specificity)}=0.200000
\]

In the 2\textsuperscript{nd} level, 145 features (removal of clinical lexicon) are used to classify factual and subjective questions. 50 features are selected.

\[
\text{mean(accuracy)}=0.853810 \quad \text{std(accuracy)}=0.050335 \\
\text{mean(error1)}=0.103333 \quad \text{std(error1)}=0.264085 \\
\text{mean(error2)}=0.160000 \quad \text{std(error2)}=0.078253 \\
\text{mean(sensitivity)}=0.896667 \quad \text{mean(specificity)}=0.840000
\]

In the 3\textsuperscript{rd} level, 145 features (removal of clinical lexicon) are used to classify opinion/advice and emotional support questions.

\[
\text{mean(accuracy)}=0.933333 \quad \text{std(accuracy)}=0.083148 \\
\text{mean(error1)}=0.160000 \quad \text{std(error1)}=0.206559 \\
\text{mean(error2)}=0.020000 \quad \text{std(error2)}=0.042164 \\
\text{mean(sensitivity)}=0.840000 \quad \text{mean(specificity)}=0.980000
\]

Together, the accuracy is in Figure 3.

\[
\text{accuracy} = (55 \times 0.8967 + 51 \times 0.7056 + 99 \times 0.8232 + 12 \times 0.1668) / 217 = 0.8
\]

And comparisons are shown in Table 2.

<table>
<thead>
<tr>
<th>Question</th>
<th>Facts</th>
<th>Opinion/Advice</th>
<th>Emotional Support</th>
<th>Companionship</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical 1</td>
<td>0.843</td>
<td>0.553</td>
<td>0.654</td>
<td>0</td>
<td>0.6447</td>
</tr>
<tr>
<td>Companion</td>
<td>1*</td>
<td>1*</td>
<td>0.84</td>
<td>0.84</td>
<td>0.6855</td>
</tr>
<tr>
<td>Factual</td>
<td>1*0.8967</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subjective</td>
<td>1*0.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opinion/Advice</td>
<td>0.84</td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emotional</td>
<td>0.84*0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Binary SVM accuracy results

Table 2. Accuracy comparisons of results
This project firstly proposed a hierarchy for an online cancer forum. Clinical features and subjective features are added into basic features with high term frequency and importance. Binary SVM is used to classify a binary tree. The method can be applied in other online health forum (different aspect of clinical features may be learned).

Acknowledgments
Raw data is provided by American Cancer Society®. (www.cancer.org)

References


