Maximal Information Coefficient for Feature Selection for Clinical Document Classification

Chen Lin
Children’s Hospital of Boston, 300 Longwood Ave., Boston, MA 02115
chen.lin@childrens.harvard.edu

Helena Canhao
Rheumatology Research Unit, Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa, Lisbon, Portugal
helenacanhao@netcabo.pt

Timothy Miller
Children’s Hospital of Boston, 300 Longwood Ave., Boston, MA 02115
Timothy.Miller@childrens.harvard.edu

Dmitriy Dligach
Children’s Hospital of Boston, 300 Longwood Ave., Boston, MA 02115
Dmitriy.Dligach@childrens.harvard.edu

Robert Martin Plenge
Brigham and Women’s Hospital, 75 Francis Street, Boston, MA 02115
RPLENGE@PARTNERS.ORG

Elizabeth Wood Karlson
Brigham and Women’s Hospital, 75 Francis Street, Boston, MA 02115
EKARLSON@PARTNERS.ORG

Guergana Savova
Children’s Hospital of Boston, 300 Longwood Ave., Boston, MA 02115
Guergana.Savova@childrens.harvard.edu

Abstract

Maximal Information Coefficient (MIC) is a novel correlation statistic that measures the association strength of linear and non-linear relationships between paired variables. We describe our first attempt in applying MIC in the clinical domain for a textual feature evaluation. The effect of MIC is compared to the Pearson Correlation-based feature selection method.

1. Introduction

Recently, a novel association statistic called the maximal information coefficient (MIC) (Reshef, Reshef et al. 2011) has been proposed to measure the correlation strength between paired variables regardless of linear or nonlinear relationships. The Pearson’s correlation coefficient $r$ captures only linear association. When associations are non-linear, the accuracy of $r$ is greatly reduced. MIC, on the other hand, can provide equitable measurements for both linear and non-linear associations. For linear relationships MIC is equal to $r$.

In our clinical document classification use case, we are evaluating the relationship between a large number of variables and gold standard labels. The equitable nature of the MIC interests us in evaluating a variety of relationships. In this extended abstract, we describe work in which we use MIC as a feature selection (FS) method to help select a subset of features that are correlated with the gold standard label. We compare its effect to a conventional correlation-based feature selection (CFS) that is built on $r$.

2. Mining Clinical Text

Our use case is the identification of large patient cohorts of treatment responders/non-responders for pharmacogenomics research. Our goal is to build an automated document-level classifier to classify Electronic Medical Records (EMR) of patients with rheumatoid arthritis (RA) into dichotomous disease activity levels, Low/Remission (LR) and Moderate/High (MH). The classifier is built upon a Natural Language Processing (NLP) System, the clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova, Masanz et al. 2010). cTAKES is a high-throughput NLP platform built on Apache’s Unstructured Information Management Architecture (UIMA 2012) engineering framework. It automatically processes clinical notes and identifies clinical named entities (NEs), including diseases/disorders, signs/symptoms, anatomical sites, procedures, and medications, and converts them into their ontological mapping code – Unified Medical Language System Concept Unique Identifier (UMLS CUI) (UMLS 2012). Through cTAKES, each EMR note can be represented as a vector of numeric features. Each feature is a binary value representing a CUI term’s occurrence or non-occurrence in a document, regardless of frequency.
Our training data includes 2,792 notes which are selected from 821 patients from the Brigham and Women’s Hospital (BWH) database. 1,472 notes are labeled as MH and 1,320 are labeled as LR. An independent set of notes from another study is used as testing data. The latter set includes 800 MH notes and 949 LR notes from 852 BWH patients.

3. Experiment Setup

Through cTAKES each clinical note is represented as 6,541 binary CUI features. The goal of our evaluation is to compare MI-based FS to CFS which is based on Pearson correlation. Unfortunately, CFS is not viable for a large feature space such as ours. To make the comparison possible, we reduce our feature space first. We use a frequency cutoff and Chi-square statistic to first reduce the number of features to 506. If a CUI feature appears only once across the training corpus, or if a feature is present in every document, it is filtered out. If the Chi-square value between a feature and the label equals zero (totally independent), this feature is filtered out. We describe the details of our work on the feature selection pipeline in a companion submission (Lin, Miller et al. 2012).

We use MIC to measure the correlation strength between each feature and the gold standard label. The histogram of 506 MIC values is plotted in Figure 1. The majority values are less than 0.02. With MIC value of 0.02 as a threshold, 54 features are selected. In comparison, we also use Pearson Correlation-based Feature Selection (CFS) (Hall 1999) with Genetic Algorithm search to evaluate these 506 features and select 169 features. A Support Vector Machine (SVM) classifier with a linear kernel is trained on both feature sets. The model performance is measured through validation on the test set. The performance was measured by $F_1$-score and Area Under Curve (AUC). $F_1$-score is the weighted mean between recall (R) and precision (P): $F_1 = \frac{2*P*R}{P+R}$, where recall is $R = TP/(TP+FN)$ and precision is $P = TP/(TP+FP)$ where TP is true positives, FP is false positives, FN is false negatives. AUC is the overall model performance.

4. Results

MIC ends up selecting much fewer features than CFS does (54 features vs. 169 features), which improves the efficiency of model training. Table 1 shows the results on the test set for both feature sets. MIC-select features achieve better performance.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>$F_1$-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIC</td>
<td>0.756</td>
<td>0.795</td>
</tr>
<tr>
<td>CFS</td>
<td>0.742</td>
<td>0.777</td>
</tr>
</tbody>
</table>

Acknowledgments

The work was funded by PharmacoGenomics Research Network 1U01GM092691-01 (PI Plenge) and Informatics for Integrating Biology and the Bedside U54LM008748 (PI Kohane).

References


Figure 1. Histogram of MIC scores of 506 CUI features.