Using community structure detection to rank annotators when ground truth is subjective

Haimonti Dutta and William Chan
The Center for Computational Learning Systems (CCLS) and Department of Computer Science
Columbia University
New York, NY 10115
{haimonti@ccis.,wc2372@columbia.edu

Abstract

Learning using labels provided by multiple annotators has attracted a lot of interest in the machine learning community. With the advent of crowdsourcing cheap, noisy labels are easy to obtain. This has raised the question of how to assess annotator quality. Prior work uses bayesian inference to estimate consensus labels and obtain annotator scores based on expertise; the key assumptions are that the ground truth is known and categories of labels are predefined. In applications where it is possible to have multiple ground truths, assessing annotator quality is challenging since the ranking of annotators’ is dependent on the choice of ground-truth. This paper describes a case-study in the context of annotating historic newspaper articles from the New York Public Library. The goal is to assign fine-grained categorization of articles labeled “editorial” by the Optical Character Recognition (OCR) software. The task is subjective since pre-defined categories are not available. To define the ground truth we use a Community Structure Detection (CSD) algorithm in a similarity graph formed between articles. The labels from the CSD algorithm provides the target function to be learned. Annotators labels are then viewed as related tasks that help learn this target function. The technique helps to provide insights into how to rank annotator performance using well known information retrieval metrics.

1 Introduction

Chronicling America\(^1\) is an initiative of the Library of Congress (LC) whose goal is to develop an online, searchable database of historically significant newspapers between 1836 and 1922. The New York Public Library (NYPL) is part of this initiative and has scanned 200,000 newspaper pages published between 1890 and 1920 from microfilm. Automatic categorization of articles from this archive is challenging – newspapers are scanned on a page-by-page basis and article level segmentation is poor or non-existent; the OCR scanning process is far from perfect and documents generated from it contains a large amount of garbled text. The OCR software provides a high level categorization of articles, but this is often not very helpful – for instance, an attempt to categorize articles in the edition of The Sun newspaper published on November 4th, 1894 resulted in 338 articles classified as editorial which is approximately 82\% of the total number of articles in this issue. There is no easy way to categorize editorial articles into sub-groups – thus articles dealing with elections, crime and public health are all labeled “editorial”. The New York Public Library employed annotators to study whether a broad level categorization of the articles can be found. Since there is no oracle present to provide “ground truth”, a key question that needs to be answered is, which categorization from the annotators (or a combination) should be relied upon?

\(^1\)http://chroniclingamerica.loc.gov/
The contribution of this paper is as follows: (1) **Establishing ground truth:** A Community Structure Detection (CSD) algorithm is run on the similarity graph formed by articles – the detected communities are marked as “ground truth” categories; this process, however, is able to generate multiple ground truths depending on the degree of similarity between articles – an user-defined parameter. **Second opinion:** Categories provided by annotators are similar to a “second” opinion and help learn the ground truth(s). This formulation of the problem helps define a baseline (with CSD) against which alternative suggestions (from annotators) can be compared. (2) Information retrieval metrics (such as correlation coefficient, normalized mutual information, and adjusted rand index) are used to provide heuristics for comparing different partitions and for measuring the performance of annotators. (3) A case-study is presented on a small subset of articles sampled from the archive in a controlled environment, using volunteer annotators.

## 2 Background

### The Data:

The historical newspaper archive contains two types of XML files: (1) **Page-Level XMLs:** For each page of a newspaper, there is an XML file that contains metadata about the page and the text in it. (2) **Issue-Level XMLs:** The issue-level XMLs provide the following information about articles: (a) *Headlines cleaned by humans* which are of much higher quality than the text produced by the OCR software. (b) *Article segmentation information.* Each newspaper article is represented as a collection of one or more text blocks and their pixel coordinates are available. This helps to determine where one article ends and the next one begins and is particularly useful when an article spans more than one page. (c) *High-level categorization* of the articles produced by the OCR software. For the case study reported in this paper, articles are sampled randomly from one newspaper (November 2nd, 1894 issue of *The Sun*) from the archive.

### Pre-processing:

We first preprocess the articles to reduce dimensionality and have clean data to learn from. For each article, a bag-of-words representation and tf-idf weights are obtained. Stop words such as “the”, “and”, etc. are removed from the set of words. Terms of length three or less and words that contain digits or repeated characters (e.g. “paaa” and “ornnn”) are also removed. After applying the above noise reduction techniques, the dimensionality of the feature space is 3210.

### Ground Truth:

To establish ground truth, a similarity graph from the articles is built and Community Structure Detection (CSD) algorithms [2] are run on it, to partition the graph into meaningful sub-categories. Each article forms a vertex and an edge exists between two vertices if the cosine similarity between the term frequency-inverse document frequency between articles exceeds a user-defined threshold. The choice of different thresholds naturally generate multiple ground truths.

### Community Structure Detection:

Community structure detection has its roots in network theory [1]. It uses properties of a graph to find communities. A guiding principle for obtaining good division of a graph into communities is to find different number of edges between communities compared to what is expected [2]. This implies that if it is observed that the number of edges between two groups is only what one would have been expected on the basis of random chance, then this is indicative of good community structure. However, if the number of edges is much smaller or larger than this expected value, something interesting may be going on. **Modularity** of a graph is defined to be the number of edges falling within communities minus the expected number in an equivalent graph with edges placed at random. A positive modularity value indicates intracommunity connections, while negative modularity indicates a lack of such connections. We use a community structure detection algorithm based on modularity maximization [3] to establish “ground truth” in all empirical analysis.

### Second Opinion:

Annotators provide noisy labels [4]. Let $y_{ij}^k$ be the label provided by the $j$-th annotator to the $i$-th instance. The labels can be categorical and it is assumed that there are $k$ such categories i.e. $y_{ij}^k$ can take values between $c_1, c_2, \ldots, c_k$. The ground truth labels are given by $y_i^{[p]}$, $1 \leq p \leq m$, assuming $m$ ground truths can be defined. The annotators help learn each ground-truth independently. If $y_{ij}^k = c_k$ when $y_i^{[p]} = c_k$, $p \in [1, m]$, then the annotators labels match with a realization of the ground truth.
3 Case Study

Six annotators were recruited to determine the number of natural categories found in a random sample of twenty-five articles from the archive. All the annotators were given the same set of articles to work with. They were asked to skim the articles first and group them into obvious and intuitive categories by focusing on the “big picture”. The defined categories had to be described in 5 - 10 words and preferably had to include words from the articles. Finally, they were interviewed with the following set of questions: (1) What was the strategy you used for coming up with the categories? (2) Were there any documents that you found difficult to assign to categories? (3) Did you find any part of the study particularly difficult or ambiguous? If so, describe the problem you faced. (4) How long did it take you to complete the study? (5) If you had the opportunity to change anything with this study, what would it be? The focus of the case study is on determining which categorization provided by the annotators can be relied upon.

4 Metrics

Different information retrieval metrics are used to compare performance of an annotator against one of the possible ground truth(s) obtained from the CSD algorithm. Assume that data set \( D \) has \( n \) instances \( O_1, O_2, \ldots, O_n \) and we want to partition it into \( K \) communities. Let \( K = \{1, 2, \ldots, \eta\} \) be the set of labels assigned by the CSD algorithm and \( C = \{1, 2, \ldots, \eta\} \) be the annotated class labels assigned to the objects in \( D \). The “ground truth” is compared to the “second opinion” using correlation coefficient, normalized mutual information and adjusted rand index.

5 Experimental Results

25 articles sampled randomly from the November 2nd, 1894 issue of The Sun newspaper form the vertices of a similarity graph and thresholded cosine similarity between the term frequency-inverse document frequency is used to determine whether an edge should exist between two vertices. By adjusting the threshold\(^2\) of cosine similarity, it is possible to create a large set of graphs with differing number of edges. The cosine similarity value in our application ranged from 0.03 to 0.5. There are a total of twenty-three graphs (after eliminating duplicates) on each of which Newman’s algorithm for community structure detection is run. Many graphs resulted in only one or two communities. The graphs with thresholds between 0.03 and 0.09 all had two or fewer communities, as was the case with thresholds 0.11, 0.30, and from 0.34 to 0.46. Although the articles can be grouped into just one or two categories, this was not a particularly interesting result. Most users would prefer a higher number of categories of articles, so the analysis was restricted to those that produced more than four communities. An upper bound on the cosine similarity is also imposed to remove any thresholds that produced graphs with isolated nodes.

<table>
<thead>
<tr>
<th>Annotator</th>
<th>Th (#Cms)</th>
<th>1(8)</th>
<th>2(13)</th>
<th>3(13)</th>
<th>4(9)</th>
<th>5(10)</th>
<th>6(13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10(9)</td>
<td>0.34</td>
<td>0.28</td>
<td>0.07</td>
<td>0.26</td>
<td>0.31</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>0.12(23)</td>
<td>-0.26</td>
<td>0.53</td>
<td>0.78</td>
<td>0.58</td>
<td>0.57</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>0.13(10)</td>
<td>0.06</td>
<td>0.23</td>
<td>0.36</td>
<td>0.18</td>
<td>0.23</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>0.14(17)</td>
<td>0.02</td>
<td>0.67</td>
<td>0.60</td>
<td>0.56</td>
<td>0.55</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>0.15(5)</td>
<td>-0.36</td>
<td>0.11</td>
<td>0.14</td>
<td>0.09</td>
<td>0.22</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>Ann Clus</td>
<td>8</td>
<td>14</td>
<td>12</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: (Left) Correlation Coefficients and (Right) Normalized Mutual Information Comparing Annotators to Community Structure Detection Results

The information retrieval metrics discussed above are applied to the communities obtained by running the CSD algorithm and the labels suggested by annotators. The results from using correlation

\(^2\)If the cosine similarity is above the threshold, an edge exists between vertices, otherwise not.
coefficient is shown in Table 1\(^3\) (Left). An interesting observation obtained from using this metric is that the first annotator who suggested 8 clusters correlates negatively with suggestions from the community structure detection algorithm which suggests either 23 (too many) or 5 (too few) clusters. When the number of clusters suggested by annotators is 9, 10 or 13 though, the correlation coefficient leads one to believe that the partition with 23 clusters obtained from the community detection algorithm is the one it resembles most. A similar phenomenon is seen when using normalized mutual information (Table 1 (Right)). This is somewhat misleading, and the metrics are not very useful in showing which annotator does better (or worse) than others.

<table>
<thead>
<tr>
<th>Th (#Cms)</th>
<th>Annotator</th>
<th>1(8)</th>
<th>2(13)</th>
<th>3(13)</th>
<th>4(9)</th>
<th>5(10)</th>
<th>6(13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10(9)</td>
<td>0.06</td>
<td>0.11</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>0.12(23)</td>
<td>0.03</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>0.13(10)</td>
<td>0.06</td>
<td>-0.06</td>
<td>-0.08</td>
<td>-0.02</td>
<td>-0.08</td>
<td>-0.03</td>
<td></td>
</tr>
<tr>
<td>0.14(17)</td>
<td>-0.04</td>
<td>-0.01</td>
<td>-0.02</td>
<td>-0.05</td>
<td>-0.02</td>
<td>-0.06</td>
<td></td>
</tr>
<tr>
<td>0.15(5)</td>
<td>0.11</td>
<td>0.09</td>
<td>0.12</td>
<td>0.07</td>
<td>0.05</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Adjusted Rand Index Comparing Annotators to Community Structure Detection Results

Adjusted Rand Index, on the other hand, provides a different picture. Table 2 shows the result. When the annotators suggested greater than 9 clusters, it does not always resemble the 23 partitions provided by the community structure detection algorithm. This is considerably in tune with what one would expect from a metric to measure the performance of annotators. Thus, we find that ARI is more suited to the task at hand than NMI.

6 Conclusion

Chronicling America is an initiative of the Library of Congress that aims to make historical newspapers available for searching on the internet. To ensure good search quality, the articles need to be categorized. A crowdsourcing case study is designed in which annotators are asked to provide high level categories. Without an unique ground truth, it is difficult to find out whom to trust and how to test the quality of the categorization. In this paper, a novel way to rank annotators by comparing them to community structures detected using an automatic modularity estimation algorithm is presented. The results from annotators are compared to the automatic labels using information retrieval metrics such as normalized mutual information, correlation coefficient and adjusted rand index. The adjusted rand index seems to present characteristics that help to rank performance of annotators.

Acknowledgments

This work is supported by funding from the National Endowment for Humanities, Grant No: NEH HD-51153-10.

References


\(^3\)The column Th (#Cms) indicates threshold (number of corresponding communities formed) in all the tables in the paper. Similarly Annotator \(xx\) (\(yy\)) indicates that annotator number \(xx\) has suggested \(yy\) sub-categories from the data.