Extending Multivariate Bernoulli and Multinomial Models for Clustering MEDLINE Records

Shanfeng Zhu\(^1\)
zhusf@kuicr.kyoto-u.ac.jp

Ichigaku Takigawa\(^1\)
takigawa@kuicr.kyoto-u.ac.jp

Shuqin Zhang\(^2\)
sqzhang@hkusua.hku.hk

Hiroshi Mamitsuka\(^1\)
mami@kuicr.kyoto-u.ac.jp

\(^1\) Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto, Japan
\(^2\) Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong

Keywords: MEDLINE, text clustering, probabilistic model

1 Introduction

Mining the biomedical text for the biological knowledge discovery and hypothesis generation has become a very active field [1]. As the largest biomedical literature database, MEDLINE indexed over 16 million citations of biomedical documents, and thus becomes the main source of biomedical text mining. As an unsupervised learning method, clustering can explore the text collections without prior knowledge, and try to group the documents into different topics, which could help us to navigate and locate the documents of interest. Two classical models, multivariate Bernoulli and multinomial model, are widely used in practice[2].

2 Method and Results

Currently each document is deemed as an integrated object by existing algorithms, whereas a biomedical scientific article usually consists of several distinct fields, such as title, abstract, keywords, main text, and reference. Since each field has a distinguished role in presenting the idea of the document, these fields have different distributions over different vocabularies. Considering the wide usage of the classic multivariate Bernoulli and multinomial models, we try to extend these two models to handle clustering MEDLINE records with multiple fields. Here we illustrate our method by extending multivariate Bernoulli model.

Probabilistic Structure:

\[
L(D) = \sum_{d \in D} \log p(d) = \sum_{d \in D} \log \left( \sum_{z \in Z} p(z)p(d|z) \right)
\]

\[
= \sum_{d \in D} \log \left( \sum_{z \in Z} \left( p(z) \prod_{c \in C} \prod_{w \in d_c} p(w|z,c)^{B_{w,dc}} (1 - p(w|z,c))^{1-B_{w,dc}} \right) \right)
\]

E-step:

\[
p(z|d) \propto p(z) p(d|z) = p(z) \prod_{c \in d_c} \left( p(w|z,c)^{B_{w,dc}} \times (1 - p(w|z,c))^{1-B_{w,dc}} \right)
\]
M-step: (with Laplacian smoothing)

\[
p(z) \propto \sum_{d \in D} p(z|d)
\]

\[
p(w|z,c) = \frac{1 + \sum_{d \in D} p(z|d) \cdot B_{wd,c}}{2 + \sum_{d \in D} p(z|d)}
\]

We used the TREC genomic track 2004 and 2005 data to evaluate the performance of proposed probabilistic clustering models in terms of clustering MEDLINE records with multiple fields. In TREC genomic track 2005 data, we focus on top 8 topics that have the largest number of definitely relevant documents (more than 100). Similarly, from TREC genomic track 2004 data, top 9 topics that have the largest number of definitely relevant documents (more than 150) have been selected to formulate the second experimental dataset, Genomics04. We used normalized mutual information (NMI) to compare the performance of classic clustering models with the extended models. The NMI value ranges from 0 to 1, where a NMI value of closing to 0 means almost random partition, and a NMI value of closing to 1 means achieving a clustering result almost identical to the true class labeling. As shown in Table 1, bert, bera, berm and berw refer to the original Bernoulli model on title, abstract, Mesh and whole text, respectively, where fiber refers to the fielding multivariate Bernoulli model. For multinomial model, similar explanation applies to mnst, mnsa, mnsm, mnsw and fnns.

<table>
<thead>
<tr>
<th>Model</th>
<th>Genomics05</th>
<th>Genomics04</th>
<th>Model</th>
<th>Genomics05</th>
<th>Genomics04</th>
</tr>
</thead>
<tbody>
<tr>
<td>bert</td>
<td>.696 ± 0.04</td>
<td>.684 ± 0.03</td>
<td>mnst</td>
<td>.751 ± 0.04</td>
<td>.765 ± 0.03</td>
</tr>
<tr>
<td>bera</td>
<td>.714 ± 0.05</td>
<td>.797 ± 0.05</td>
<td>mnsa</td>
<td><strong>.779 ± 0.04</strong></td>
<td>.826 ± 0.05</td>
</tr>
<tr>
<td>berm</td>
<td>.663 ± 0.03</td>
<td>.787 ± 0.04</td>
<td>mnsm</td>
<td>.599 ± 0.04</td>
<td>.763 ± 0.03</td>
</tr>
<tr>
<td>berw</td>
<td>.704 ± 0.04</td>
<td>.801 ± 0.04</td>
<td>mnsw</td>
<td>.768 ± 0.04</td>
<td><strong>.838 ± 0.04</strong></td>
</tr>
<tr>
<td>fiber</td>
<td><strong>.741 ± 0.04</strong></td>
<td><strong>.824 ± 0.04</strong></td>
<td>fnns</td>
<td>.770 ± 0.04</td>
<td>.837 ± 0.04</td>
</tr>
</tbody>
</table>

3 Discussion

On these two datasets, we find that fielding multivariate Bernoulli model outperforms original model significantly, and fielding multinomial model has similar similar performance with original model. We will examine these models on more datasets in the future.

References
