A re-examination of text categorization methods

Barış Geçer
M. Ozan Karsavuran
Problem Definition

Statistical significance test on five text categorization methods with a skewed category distribution:

- SVM
- kNN
- NNet
- LLSF (Linear Least Squares Fit)
- NB
Motivation

- Cross method comparison (NNet vs SVM ?)
- Robustness on skewed category distribution
  - In real life, they are extremely non-uniform
- Effectiveness of each method as a function of rareness of categories
  - Single score: accuracy, error rate, F1 measure
    - can be dominated by common classes
  - Multi score: Micro-averaging, macro-averaging
Contributions

- Comparison of five methods on the new benchmark corpus
- Variety of statistical significance analysis and suggestion to combine them
- Performances as a function of category frequency
  - i.e. skewed category distribution
Benchmark Corpus

- 82% of categories have less than 100 instances
- 33% have less than 10 instances

Figure 1: Category distribution in Reuters-21578 AptoMod
Performance Measures

- Macro-averaging: F1 measure computed for each category individually then averaged
- Micro-averaging: F1 measure computed globally
- Providing both kinds of scores is more informative than providing either alone
- Error
Figure 3: The decision line with the maximal margin. The data points on the dashed lines are the Support Vectors.
LLSF and kNN

- Although they differ statistically, they had similar performance in the authors’ previous studies
- Yet, their robustness in dealing with rare categories is unknown.
Neural Network (NNet)

- Different Networks
- Separate NNet per category
- Training cost is high
  - One NNet for all 90 categories
  - one hidden layer
Naive Bayes (NB)

- Use joint probabilities of words and categories
  - assume words are independent
Significance Tests

- s-test and p-test at micro level
- others at macro level

- Micro sign test (s-test)
- Macro sign test (S-test)
- Macro t-test (T-test)
- Macro t-test after rank transformation
- Comparing proportions (p-test)
Significance Tests

- S-test: robust for reducing the influence of outliers but risks being insensitive
- T-test: could be overly sensitive when F scores are unstable
- T’-test: less sensitive to outliers but more sensitive than sign tests
Significance Tests

- None of them is “perfect”
  - for skewed category distribution
- So use them jointly
Evaluation

Different size of features that optimize the F score for each classifier
Comparison

Table 1: Performance summary of classifiers

<table>
<thead>
<tr>
<th>method</th>
<th>miR</th>
<th>miP</th>
<th>miF1</th>
<th>maF1</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>.8120</td>
<td>.9137</td>
<td>.8599</td>
<td>.5251</td>
<td>.00365</td>
</tr>
<tr>
<td>KNN</td>
<td>.8339</td>
<td>.8807</td>
<td>.8567</td>
<td>.5242</td>
<td>.00385</td>
</tr>
<tr>
<td>LSF</td>
<td>.8507</td>
<td>.8489</td>
<td>.8498</td>
<td>.5008</td>
<td>.00414</td>
</tr>
<tr>
<td>NNet</td>
<td>.7842</td>
<td>.8785</td>
<td>.8287</td>
<td>.3765</td>
<td>.00447</td>
</tr>
<tr>
<td>NB</td>
<td>.7688</td>
<td>.8245</td>
<td>.7956</td>
<td>.3886</td>
<td>.00544</td>
</tr>
</tbody>
</table>

miR = micro-avg recall; miP = micro-avg prec.; miF1 = micro-avg F1; maF1 = macro-avg F1.

miF1 of SVM is lower than Joachims but not significant
Comparison

<table>
<thead>
<tr>
<th>method</th>
<th>miR</th>
<th>miP</th>
<th>miF1</th>
<th>maF1</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>.8120</td>
<td>.9137</td>
<td>.8599</td>
<td>.5251</td>
<td>.00365</td>
</tr>
<tr>
<td>KNN</td>
<td>.8339</td>
<td>.8807</td>
<td>.8567</td>
<td>.5242</td>
<td>.00385</td>
</tr>
<tr>
<td>LSF</td>
<td>.8507</td>
<td>.8489</td>
<td>.8498</td>
<td>.5008</td>
<td>.00414</td>
</tr>
<tr>
<td>NNet</td>
<td>.7842</td>
<td>.8785</td>
<td>.8287</td>
<td>.3765</td>
<td>.00447</td>
</tr>
<tr>
<td>NB</td>
<td>.7688</td>
<td>.8245</td>
<td>.7956</td>
<td>.3886</td>
<td>.00544</td>
</tr>
</tbody>
</table>

miR = micro-avg recall; miP = micro-avg prec.; miF1 = micro-avg F1; maF1 = macro-avg F1.

miF1 of kNN is higher than Joachims, simplified kNN is similar:
it is neither optimal nor necessary
Comparison

Table 1: Performance summary of classifiers

<table>
<thead>
<tr>
<th>method</th>
<th>miR</th>
<th>miP</th>
<th>miF1</th>
<th>maF1</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>.8120</td>
<td>.9137</td>
<td>.8599</td>
<td>.5251</td>
<td>.00365</td>
</tr>
<tr>
<td>KNN</td>
<td>.8339</td>
<td>.8807</td>
<td>.8567</td>
<td>.5242</td>
<td>.00385</td>
</tr>
<tr>
<td>LSF</td>
<td>.8507</td>
<td>.8489</td>
<td>.8498</td>
<td>.5008</td>
<td>.00414</td>
</tr>
<tr>
<td>NNet</td>
<td>.7842</td>
<td>.8785</td>
<td>.8287</td>
<td>.3765</td>
<td>.00447</td>
</tr>
<tr>
<td>NB</td>
<td>.7688</td>
<td>.8245</td>
<td>.7956</td>
<td>.3886</td>
<td>.00544</td>
</tr>
</tbody>
</table>

miR = micro-avg recall; miP = micro-avg prec.; miF1 = micro-avg F1; maF1 = macro-avg F1.

miF1 of NB is higher,
multinomial mixture vs multivariate Bernoulli
# Comparison

## Table 2: Statistical significance test results

<table>
<thead>
<tr>
<th>sysA</th>
<th>sysB</th>
<th>s-test</th>
<th>S-test</th>
<th>T-test</th>
<th>T’-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>kNN</td>
<td>&gt;</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>SVM</td>
<td>LLSF</td>
<td>&gt;</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>kNN</td>
<td>LLSF</td>
<td>&gt;</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>SVM</td>
<td>NNet</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>kNN</td>
<td>NNet</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>LLSF</td>
<td>NNet</td>
<td>?</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>NB</td>
<td>kNN</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>NB</td>
<td>LLSF</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>NB</td>
<td>SVM</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
<tr>
<td>NB</td>
<td>NNet</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
<td>&lt;</td>
</tr>
</tbody>
</table>
Comparison

- micro level:
  ○ SVM > kNN >> {LLSF, NNet} >> NB
- macro level:
  ○ {SVM, kNN, LLSF} >> {NB, NNet}
- micro: dominated by common categories
- macro: dominated by rare categories
  ○ complementary
Conclusions

- Significance analysis on five well-known classifiers
- micro-level, macro-level and joint for cross comparison
- significance depends on performance measure
Thank you!

Questions & Answers