A Generic Classifier-Ensemble Approach for Biomedical Named Entity Recognition

Zhihua Liao
MFLETC, Hunan Normal University
Changsha, China

PAKDD2012 Kuala Lumpur, Malaysia
Outline

- Motivation
- Current Methods
- Our Model
- Experimental Results
- Conclusion
Motivation

- Single classifier's limitations
- Combined-classifier has more space to improve
- Combined-classifier method has much strong theoretical fundamentals
  - Diversity measure
  - Majority voting
- Many applications based on classifier-ensemble approach, e.g. text classification, bioinformatics mining

PAKDD2012 Kuala Lumpur, Malaysia
Current Methods

- Traditional individual classifier
  - CRF
  - HMM
  - MEMM
  - SVM
  - ......

- Classifier-ensemble method
  - Zhang & Yang [2008]
  - Yang et al. [2010]
  - Thomas G, Dietterich [2000]
  - Dymitr & Bogdan [2001]
  - Nazife et al. [2007, 2008, 2009]
Our Model: the main idea

- A classifier-ensemble system usually consists of two major components
  - The classifiers forming the ensemble members
  - The combination scheme
- SVM is used as the basic classifier
- The genetic algorithm is used to search the optimal solution of weighted classifier combination
  - Chromosome
  - Fitness

Both are two core elements in GA

PAKDD2012 Kuala Lumpur, Malaysia
Feature Set and SVM Based Classifier

Feature set
- words
- orthographic
- prefix and suffix
- lexical
- preceding class
- surface word

SVM classifier - the parameters of Yamcha
- kernel
- degree of kernel
- direction of parsing
- windows position
- multi-class
Genetic Classifier-Ensemble-I

- The basic idea: different classes in each classifier differ with contributing degrees of prediction classes.
- The definition of the combined score of a particular class:
- The fitness as follows:

\[
\begin{align*}
    f_n(c_1) &= \max(f(c_1), f(c_2), \ldots, f(c_n)) \\
    f(c_i) &= \sum_{m=1}^{M} F_m \ast w(m, i)
\end{align*}
\]

Fig. 1. Genetic Classifier-Ensemble-I
The basic idea: different classifiers have different contributing degrees of prediction of classes.

The combined score of a given class can be redefined as:

The fitness NOT changes

\[ f(c_i) = \sum_{m=1}^{M} F_m \times w_m \]

**Fig. 2. Genetic Classifier-Ensemble-II**
The idea: not only contributing degrees of prediction classes among different classes in the same classifier are different, but also contributing degrees of prediction classes among different classifiers differ.

The chromosome in this method is made up of the chromosome in genetic classifier-ensemble-I and the one in genetic classifier-ensemble-II.

The combined score of a given class is determined as:

$$f(c_i) = \sum_{m=1}^{M} F_m \ast w(m,i) \ast w_m$$
A workflow of the generic genetic classifier-ensemble algorithm

Fig. 4. The flow of the proposed generic genetic classifier-ensemble algorithm.

PAKDD2012 Kuala Lumpur, Malaysia
Our system’s framework

Fig. 5. Overall system architecture.

PAKDD2012 Kuala Lumpur, Malaysia
Experimental Results

- Evaluation criteria
  - Precision
  - Recall
  - F1

- Dataset (GENIA 3.02 corpus) used for evaluation

<table>
<thead>
<tr>
<th>Types</th>
<th>Train data</th>
<th>Test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>9534</td>
<td>1056</td>
</tr>
<tr>
<td>RNA</td>
<td>951</td>
<td>118</td>
</tr>
<tr>
<td>Cell_line</td>
<td>3830</td>
<td>500</td>
</tr>
<tr>
<td>Cell_type</td>
<td>6718</td>
<td>1921</td>
</tr>
<tr>
<td>Protein</td>
<td>30269</td>
<td>5067</td>
</tr>
<tr>
<td>Total</td>
<td>51302</td>
<td>8662</td>
</tr>
</tbody>
</table>
Experimental Results

In this table, the genetic classifier-ensemble-III gets the better results compared with the genetic classifier-ensemble-I and genetic classifier-ensemble-II.

Table 5. The performances of different biomedical named entities on three genetic classifier-ensemble schemes.

<table>
<thead>
<tr>
<th>Types</th>
<th>Genetic Classifier-ensemble-I</th>
<th>Genetic Classifier-ensemble-II</th>
<th>Genetic Classifier-ensemble-III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision Recall F-score</td>
<td>Precision Recall F-score</td>
<td>Precision Recall F-score</td>
</tr>
<tr>
<td>DNA</td>
<td>73.54 74.25 72.92</td>
<td>70.68 70.98 70.76</td>
<td>74.65 75.59 75.21</td>
</tr>
<tr>
<td>RNA</td>
<td>74.33 75.85 75.22</td>
<td>71.15 70.25 70.46</td>
<td>75.88 76.79 76.42</td>
</tr>
<tr>
<td>Cell_line</td>
<td>72.50 71.56 72.12</td>
<td>68.25 67.20 67.82</td>
<td>74.60 73.82 74.36</td>
</tr>
<tr>
<td>Cell_type</td>
<td>73.15 72.87 72.04</td>
<td>69.62 72.58 70.37</td>
<td>74.85 75.39 75.06</td>
</tr>
<tr>
<td>Protein</td>
<td>83.36 76.58 79.65</td>
<td>80.56 71.25 75.86</td>
<td>84.58 80.06 83.57</td>
</tr>
<tr>
<td>Total</td>
<td>74.33 73.52 73.86</td>
<td>71.28 71.02 71.16</td>
<td>75.65 78.52 77.85</td>
</tr>
</tbody>
</table>
Compared with the individual best SVM-classifier and SVM-classifier ensemble, our method outperforms them.

Table 6. The comparison with individual best SVM classifier and Vote-based SVM-classifier selection for bioNER task.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single best SVM-classifier[20, 21]</td>
<td>69.40</td>
<td>70.60</td>
<td>69.99</td>
</tr>
<tr>
<td>Vote-based SVM-classifier selection[20, 21]</td>
<td>71.74</td>
<td>73.76</td>
<td>72.74</td>
</tr>
<tr>
<td>Genetic classifier-ensemble-III</td>
<td>75.65</td>
<td>78.52</td>
<td>77.85</td>
</tr>
</tbody>
</table>

It shows that the best result of our experiment outperforms that of other individual classifier algorithms.

Table 7. The comparison with other different individual classifier algorithms on bioNER task.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhou and Su[1]</td>
<td>69.42</td>
<td>75.99</td>
<td>72.55</td>
</tr>
<tr>
<td>Finkel et al.[2]</td>
<td>68.56</td>
<td>71.62</td>
<td>70.06</td>
</tr>
<tr>
<td>Settles[3]</td>
<td>69.30</td>
<td>70.30</td>
<td>69.80</td>
</tr>
<tr>
<td>Song et al.[4]</td>
<td>64.80</td>
<td>67.80</td>
<td>66.30</td>
</tr>
<tr>
<td>Zhao[5]</td>
<td>61.00</td>
<td>69.10</td>
<td>64.80</td>
</tr>
<tr>
<td>Genetic classifier-ensemble-III</td>
<td>75.65</td>
<td>78.52</td>
<td>77.85</td>
</tr>
</tbody>
</table>
Conclusion

- We propose a generic genetic classifier-ensemble approach for biomedical named entity recognition.
- The individual classifier - SVM is adopted as SVM-classifier committee.
- A multi-objective genetic algorithm is employed as the classifier selector to facilitate the ensemble classifier to improve the overall sample classification accuracy.
- This method can extend easily to other applications of text classification.
Thank you!

Goodbye