GA based Feature Selection for Nursing-care Freestyle Text Classification

Manabu Nii1,2, Shigeru Ando1, Yutaka Takahashi1
1Graduate School of Engineering, University of Hyogo
Shosha 2167, Himeji, Hyogo, Japan
2WPI Immunology Frontier Research Center, Osaka University
Atsuko Uchinuno, Reiko Sakashita
College of Nursing Art & Science, University of Hyogo
Kitaouji-cho 13-71, Akashi, Hyogo, Japan

Abstract—The nursing care quality improvement is very important in the medical field. Currently, nursing-care freestyle texts (nursing-care data) are collected from many hospitals in Japan by using Web applications. The collected nursing-care data are stored into the database. Some nursing-care experts evaluate the collected data to improve nursing care quality. In order to evaluate the nursing-care data, experts need to read all freestyle texts carefully. However, it is a hard task for each expert to evaluate the data because of huge number of nursing-care data in the database. For reducing workloads to evaluate nursing-care data, we have proposed a support vector machine (SVM) based classification system. In this paper, in order to improve the classification performance, we propose a genetic algorithm (GA) based feature selection method for generating numerical data from collected nursing-care texts. First, we extract nouns and verbs from nursing-care texts using the morphological analysis software “MeCab” and store the extracted terms into a “term list”. Some combinations of terms in the term list are selected by GA with two objectives; (1) maximization of the number of correctly classified texts and (2) minimization of the number of selected terms. And then, we classify the nursing-care numerical data using the SVM. From computer simulation results, we show the effectiveness of our proposed method.

I. INTRODUCTION

The nursing care quality improvement is very important in the medical field. The nursing-care freestyle texts in this paper are Japanese texts written by nurses which consist of answers for questions about nursing-care. The nursing-care data are collected via Web applications from many hospitals in Japan. The collected data are stored into the database. The nursing-care experts evaluate the collected data to improve nursing-care quality.

Currently, the collected data are evaluated by experts reading all texts carefully. It is difficult, however, for experts to evaluate the data because there is huge number of nursing-care data in the database. In order to reduce workloads evaluating of nursing-care data, we have been proposed a fuzzy classification system [1], a neural network based classification system [2], and a SVM based system [3], [4]. The SVM based classification system proposed in [3], [4] had higher generalization ability than the other systems.

To classify Japanese texts written by many nurses, we perform the natural language processing for the texts. The first step of the natural language processing is morphological analysis. Because Japanese texts have no space between morphemes (i.e., words), we need to decompose texts into several morphemes. We can use some morphological analysis tools such as “ChaSen” [5]–[8], “MeCab” [9], “JUMAN” [10], etc.

In this paper, in order to improve the classification performance, we propose a genetic algorithm (GA) based feature selection method for generating numerical data from collected nursing-care texts. First, we extract nouns and verbs from nursing-care texts using the morphological analysis software “MeCab” and store the extracted terms into a “term list”. Some combinations of terms in the term list are selected by GA with the following two objectives; (1) maximization of the number of correctly classified texts and (2) minimization of the number of selected terms. And then, we classify the nursing-care numerical data using the SVM. From computer simulations, we show the effectiveness of our proposed system using the 10-fold cross validation method.

II. NURSING-CARE DATA

Currently, nursing-care texts written by nurses are collected via Web applications and stored into the database. The stored data have several question parts and the corresponding answer parts. The collected text data include many types of answers such as long, short, non-entered one, etc. This is because multiple nurses write texts about their own patients. Many kinds of terms in the field of nursing-care are also used in these texts. These terms depend on each nurse because nursing treatment that each nurse performs to a patient varies according to the patient’s disease and condition. Some terms in general use are also utilized as technical terms in the nursing-care domain. Moreover, either the technical or the general term is changed by the meaning of such a term in view of the context. Nursing-care experts can evaluate such texts correctly by interpreting their contents and syntax. Figure 1 shows an example of nursing-care data. Generally, we can see that good texts which were classified by experts are long, while bad ones are short from Fig. 1.

Today, experts evaluate all collected nursing-care texts manually. Therefore experts have heavy workloads for reading and classifying all texts and they wish to reduce these works. We have developed the SVM based system [3], [4]. Although the SVM based system has good classification performance,
we would like to improve it more. In order to improve the classification performance, we propose a feature extraction method which chooses some important words in the nursing-care texts.

III. FEATURE EXTRACTION FROM NURSING-CARE DATA

To classify nursing-care data, first we decompose each text into morphemes using morphological analysis software “MeCab.” The morpheme is the minimum unit which grammatically has the meaning. Figure 2 shows an example of a decomposed Japanese text. In Fig. 2, each morpheme was separated by slash mark. After decomposing texts, nouns and verbs are recorded in a list called “term list.”

Generally, the vector space model is used for classifying texts. In the vector space model, the \( p \)-th text is represented by an \( n \)-dimensional vector \( x \) as follows:

\[
x_p = (x_{p1}, x_{p2}, \ldots, x_{pn}), \quad p = 1, 2, \ldots, N_{\text{txts}},
\]

where, \( p \) is an index, \( N_{\text{txts}} \) is the number of all texts. As each attribute value, the \( tf-idf \) weights have often been used. For classifying nursing-care texts, we have defined the following vector model [3] instead of Eq.(1).

\[
x_p = (N_{\text{morph}, p}, N_{\text{term}, p}, tfidf_{\text{sum}, p}, b_1, \ldots, b_m),
\]

where, \( N_{\text{morph}, p} \) is the number of morphemes, \( N_{\text{term}, p} \) is the number of kinds of terms defined in the term list, and \( b_j, \ j = 1, 2, \ldots, m \) is the existence of terms in the term list (‘1’ means that the term appears in the text, ‘0’ means not). \( tfidf_{\text{sum}, p} \) is the summation of \( tf-idf \) weights as follows.

\[
 tfidf_{\text{sum}, p} = \sum_{t_i \in p} tfidf_{t_i, p},
\]

where \( t_i \) is a term in the term list. That is, a term \( t_i, \ i = 1, 2, \ldots, M \) is defined in the term list that contains nouns and verbs which are extracted from the nursing-care texts.

The \( tf-idf \) weight is an importance measure of a word in a document. The term frequency \( (tf) \) is the number of times that a term \( t_i \) appears in the given document (see Eq.(4)).

\[
 tf_{t_i, p} = \frac{N_{t_i, p}}{N_{\text{max}, p}},
\]

where, \( N_{t_i, p} \) is the number of times that a term \( t_i \) appears in the \( p \)-th text, and \( N_{\text{max}, p} \) is the maximum value of \( N_{t_i, p} \) in the \( p \)-th text. The document frequency \( (df) \) is the number of documents that contain the term \( t_i \). We can have the inverse document frequency \( (idf) \) by taking the logarithm as follows:

\[
 idf_{t_i} = \log \frac{N_{\text{doc}}}{df_{t_i}}.
\]

Then the \( tf-idf \) weight is calculated as,

\[
 tfidf_{t_i, p} = tf_{t_i, p} \cdot idf_{t_i}.
\]

757
IV. FORMER WORK

In [4], the following procedures are performed in order to select the terms which affect classification.

**[Former method in [4]]**

Step 1: The nursing-care data set is divided into the subset for every class.

Step 2: The frequency vector \( r_{ti} \) of appearance of each term \( t_i \) in the term list is examined for each subset. For four-class problems, we have four-dimensional vectors,

\[
\mathbf{r}_{ti} = (r_{Class1 i}, r_{Class2 i}, r_{Class3 i}, r_{Class4 i}),
\]

where \( r_{Class i} \) is normalized in \([0, 1]\) for each subset.

Step 3: \( m \) terms which appear well only to one class are chosen. First, calculate \( \text{diff}_{ti} \) as,

\[
\text{diff}_{ti} = r_{ti}^{\text{max}} - r_{ti}^{2\text{nd max}}
\]

where \( r_{ti}^{\text{max}} \) is the maximum value in the elements of \( r_{ti} \). Then, we choose \( m \) terms which have larger value of \( \text{diff}_{ti} \). That is, the term \( t_i \) which has larger value of \( \text{diff}_{ti} \) often appears into a single class.

Therefore, we can obtain \( x_p \) in Eq.(2).

V. FEATURE SELECTION USING GA

In [3], we have already shown that the classification performance is better when the number of terms (i.e., \( b_j \) in Eq.(2)) is selected by human experts, than when all terms in the term list are used. We have also been proposed an automatically selection method of terms in [4]. For selecting terms, we propose a GA-based method for considering some combinations of terms. The GA-based method can select some combinations of terms with two objectives; (1) maximization of the number of correctly classified texts and (2) minimization of the number of selected terms.

**A. String definition**

The binary strings are used. The first three features in Eq.(2) are not selected through GA. We focus on the rest of the features (i.e., \( b_1, \ldots, b_m \)) to select by GA. That is, the following string definition is used.

\[
s = (b_1, \ldots, b_l, \ldots, b_M),
\]

where \( M \) is the number of all terms in the term list, that is, the string length is equal to \( M \).

**B. Evaluation**

To satisfy two objectives (i.e., maximization of the number of correctly classified texts and minimization of the number of selected terms), we use the weighted sum method. Therefore, the objective function is,

\[
\text{Maximize } f(s) = w_1 \cdot NCP(s) - w_2 \cdot NST(s),
\]

where \( NCP(s) \) is the number of correctly classified patterns by SVM based system and \( NST(s) \) is the number of selected terms. \( w_1 \) and \( w_2 \) are weights defined by \([0, 1]\).

**C. Flow of our proposed method**

In this subsection, we summarize our proposed method. Figure 3 shows a flowchart of our method.

**[Proposed method]**

Step 1: Initialization. A population is randomly initialized. That is, a random set of terms is generated.

Step 2: Evaluation. Input vectors (see Eq.(2)) for SVM based classification system are generated. Then, each individual is evaluated by using 10-fold cross validation method. If the termination condition is not satisfied, then go to the next step.

Step 3: Crossover. Two-point crossover is used for reproduction of the next generation.

Step 4: Mutation. An individual is selected with the probability \( p_{\text{mutation}} \). Then a single bit of the selected individual is changed to the other state (i.e., 0→1 or 1→0). Thereafter go to step 2.
VI. COMPUTER SIMULATIONS

Support vector machines [11]–[13] were used for classifying nursing-care texts as well as our previous works. In this paper, we use the “LIBSVM” [14] which is one of SVM software. The LIBSVM is a popular implementation of the SVM algorithms. In the LIBSVM, the “one-against-one” approach has been adopted for multi-class classification.

We have already compared the classification performance using both sigmoid and RBF kernels in [3]. As a result, the classification performance was higher when RBF kernels were used. Therefore, in this paper, the RBF kernel is used for SVMs in computer simulations. And we used 10-fold cross validation method for evaluating the classification performance on the testing data set.

The details of the nursing-care texts and parameters of GA using in this section were shown in Table I and II. All nursing-care data sets have already been classified into four classes by nursing-care experts.

First, Fig.4 shows selected terms by human users. In this case, 30 terms were selected and we obtain 72.7% classification rate. However, for each data set, we had to choose some words manually which were considered to be useful for classification.

Next, Fig.5 shows classification results with several m values when using our former work in [4]. We can see from this figure that 76.8% texts were correctly classified with m = 250 (i.e., when 250 terms were selected from the term list). With this method, some words which are useful for classification can be selected automatically. However, combinations of terms were not considered.

We examined our proposed method with the same exper-

---

TABLE I
NURSING-CARE DATA SETS USING IN THIS PAPER.

<table>
<thead>
<tr>
<th>Data name</th>
<th>P131</th>
<th>P222</th>
<th>P425</th>
<th>P613</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>400</td>
<td>371</td>
<td>370</td>
<td>315</td>
</tr>
<tr>
<td># of Class 1</td>
<td>7</td>
<td>102</td>
<td>13</td>
<td>34</td>
</tr>
<tr>
<td># of Class 2</td>
<td>0</td>
<td>33</td>
<td>105</td>
<td>0</td>
</tr>
<tr>
<td># of Class 3</td>
<td>259</td>
<td>115</td>
<td>18</td>
<td>200</td>
</tr>
<tr>
<td># of Class 4</td>
<td>154</td>
<td>121</td>
<td>234</td>
<td>81</td>
</tr>
</tbody>
</table>

TABLE II
GA PARAMETERS.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>100</td>
</tr>
<tr>
<td>Crossover rate</td>
<td>1.0</td>
</tr>
<tr>
<td>Mutation rate</td>
<td>0.05</td>
</tr>
<tr>
<td>w1 : w2</td>
<td>1000 : 1</td>
</tr>
<tr>
<td>Generations</td>
<td>1000</td>
</tr>
</tbody>
</table>
Fig. 6. Classification rate and the number of selected terms for each generation.

VII. CONCLUSION

Conventionally, the nursing-care data were evaluated by the experts in the nursing-care field. It is difficult, however, for experts to evaluate all nursing-care data because the data are collected via Web applications from many hospitals in Japan. Therefore, we need to develop a classification system which helps experts to evaluate the collected nursing-care data.

In this paper, we proposed a GA-based feature extraction method for the nursing-care data classification system. In our former works, we had to choose some words manually which were considered to be useful for classification. With the feature selection method in [3], [4], some words which are useful for classification can be selected automatically. However, combinations of terms were not considered. Since our proposed method in this paper is based on GA, the combinations of terms in the term list are considered.
In computational experiments, 10-fold cross validation method was used for evaluating the classification performance on the testing data. From computer simulation results, we can see that the proposed GA-based feature selection method in this paper can raise classification performance.

REFERENCES


