2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
July 6-11, 2014, Beijing, China

Soft Class Decision for Nursing-care Text Classification using a K-Nearest Neighbor based System

Manabu Nii, Kazunobu Takahama, Atsuko Uchinuno, and Reiko Sakashita

***Abstract*— In the aging society such as Japan, it is very important to improve the quality of nursing-care for keeping our quality of life. Our final goal is to develop a computer aided evaluation system to improve the quality of nursing-care. For evaluating the quality of actual nursing, we have been collecting texts that are written by nurses using our Web based system. In our previous works, a SVM based classification system has been developed to classify such nursing-care texts, and a dependency relation based feature vector definition has also been proposed. The training data are pre-classified texts by a few nursing-care experts. Some texts in the training data are similar but classified into different classes. To classify the nursing-care texts with high accuracy, we need to tackle such ambiguous class labels in the training data.**

In this paper, we propose a k-nearest neighbor based classification system which can classify into classes with certainty grade.

I. Introduction

Since Japan is one of the most aging countries, many people will be patients and receive nursing. Although the number of patients increases, the number of nurses does not increase or may decrease because working population will be decreasing. Therefore, there will be many patients who are cared by a single nurse in near future. Improving the quality of nursing-care, we maintain our quality of life.

The nursing-care quality is evaluated by several aspects. One aspect is availability of facilities of hospitals and clinics. Actual nursing-care process evaluation is very difficult for evaluating by such numerical criteria which can evaluate a part of nursing quality.

To evaluate actual nursing in Japan, assessment criteria have been proposed. The assessment criteria have six domains; (1) understanding individuality of patients, (2) patient empowerment, (3) family care, (4) direct care, (5) medical team coordination, and (6) incident prevention. These criteria have been constructed for observation based evaluation by nursing-care experts. The criteria for evaluating nursing-care process are not numerical, but subjective. Although observation is the best way for evaluating actual nursing-care subjectively, it is hard to realize if a few experts evaluate every nurse in Japan. For solving this issue, a web-based nursing-care data collection system has been proposed and developing. In the web-based nursing-care data collection system, questions that are from the criteria of six domains of the nursing-care process are presented to nurses. Nurses report answers which are called nursing-care texts. The nursing-care texts are freestyle Japanese texts as answers for the presented questions. In order to evaluate actual nursing, nurses have to report their answer texts based on their actual nursing-care or experiences. The nursing-care experts review the collected texts instead of observing. The experts can evaluate actual nursing-care process by reviewing texts. After reviewing, the experts make some recommendations that consist of several improvements of nursing.

By using the web-based system, actual nursing-care process can be evaluated without observation. However, the issue of reviewing many texts by a few experts still remains. Our purpose is to develop a computer aided evaluation system. In our previous works [1]–[10], a SVM based classification system has been developed. The proposed SVM based system can classify nursing-care texts with high classification performance for some kind of datasets and slightly low performance for the other datasets. We should develop a classification system with high performance for all datasets.

In some cases, texts whose appearance of written words is similar have different class labels. Generally, we think that similar texts, which have similar appearance of written words, should be classified into the same class. However, in some cases, a slight difference of appearance has large difference of meanings. Therefore, we should consider to correctly classify such texts. Our purpose is to construct a computer aided evaluation system for the nursing-care texts. In the computer aided evaluation system, the classification results are presented to the nursing-care experts. The nursing-care experts make a recommendation based on the classification results. Let us assume that our computer aided evaluation system can provide some candidates of classification to experts when the system can not classify an evaluating text into an appropriate class. Such system is better than a system which classifies a text into a single class forcibly.

In this paper, a hybrid system which consists of a k-nearest neighbor based system and the SVM based system is proposed. Our proposed k-nearest neighbor based system can classify a text into not only a single class but also some candidate classes. Texts which can not be classified by the k-nearest neighbor based system are classified by the SVM based system that has been proposed in [10].

Manabu Nii and Kazunobu Takahama are with Graduate School of Engineering, University of Hyogo, Himeji, Hyogo, Japan (email: nii@eng.u-hyogo.ac.jp).

Manabu Nii is also with WPI Immunology Frontier Research Center, Osaka University, Suita, Osaka, Japan.

Atsuko Uchinuno, and Reiko Sakashita are with College of Nursing Art and Science, University of Hyogo, Akashi, Hyogo, Japan.

This work was supported by JSPS KAKENHI Grant Number 25463332.

978-1-4799-2072-3/14/$31.00 ©2014 IEEE

II. NURSING-CARE TEXTS

We have 12 kinds of text sets of nursing-care texts for each year's data which were collected from 2007 to 2009.

Nurses report their answers for the questions by writing freestyle Japanese texts, which represent their own actual nursing process. If nurses can not answer questions because they are inexperienced in the questions' situation, the nurses have to report the nearest situation or leave empty their answers. In such situation, answering without any experience is not allowed and such nursing-care text is not appropriate answer for the question. However, the collected nursing-care texts include such inappropriate texts. The nursing-care texts consist of wide variety contents, which include their own patients' physical condition and mental state, technicality of their own nursing, and some kinds of shortened expression of their local use, etc.

The class distributions of each dataset are shown in the table I. The column Q_l means the name of each subset, and C0, C1, C2, and, C3 are class labels. C0 means "bad nursing-care" or "an inappropriate answer for this question". C1, C2, and, C3 mean "slightly good", "good", and, "very good", respectively. We can see from the table that the class distribution for each year on the same dataset is varied. One reason is that nurses who joined our research were partly different every year. In actual use of our system, such situation needs to be considered.

III. DEPENDENCY RELATION BASED FEATURE VECTOR DEFINITION

The common preprocessing for handling Japanese texts is to decompose texts into words. Because Japanese texts are not separated by white space, every word needs to be decomposed by morphological analysis. We adopted the "MeCab [11]" because it is one of the morphological analysis software.

Every text is decomposed by MeCab into words. Then, all words are extracted and registered to a term list $TL^{Q_l}_{year}$. The term list $TL^{Q_l}_{year}$ is generated for each question Q_l every year.

The experts not only check the existence of words, but also consider the construction of a sentence when they evaluate nursing-care texts. Dependency relation between words is a kind of such constructions. In order to analyze the dependency relation of Japanese, we selected the dependency analysis software named "Cabocha". Cabocha is one of the Japanese dependency analysis software proposed in [12] and achieved about 90% of accuracy for parsing Japanese dependency. An example result of parsed dependency relations by Cabocha is shown in Fig. 1. This nursing-care text says "While I used a painkiller which was specified by the doctor, I think that my patient's pain was not controlled. So, I reported my patient's condition to the doctor, and asked to change the painkiller to the other one." Each line means a phrase. Slash mark between words is a separator of words. In Cabocha, MeCab is used for finding words.

TABLE I
DETAILS OF CLASS DISTRIBUTION OF THE 2007, 2008, AND 2009 NURSING-CARE DATASETS

2007

Q_l	C0	C1	C2	C3	Total
P123	78	0	0	459	537
P131	13	54	187	245	499
P132	199	128	99	12	438
P212	145	0	94	253	492
P213	40	101	209	158	508
P221	30	98	172	182	482
P222	111	13	66	264	454
P322	69	179	4	188	440
P411	76	0	0	315	391
P423	46	3	18	352	419
P425	25	135	43	251	454
P431	51	0	0	394	445

2008

Q_l	C0	C1	C2	C3	Total
P123	61	383	0	136	580
P131	16	293	88	128	525
P132	86	227	88	7	408
P212	166	0	54	272	492
P213	51	254	156	58	519
P221	21	97	212	193	523
P222	118	20	114	241	493
P322	39	242	4	125	410
P411	105	0	0	273	378
P423	48	23	182	149	402
P425	54	121	129	85	389
P431	115	0	0	320	435

2009

Q_l	C0	C1	C2	C3	Total
P123	42	401	0	384	827
P131	13	67	276	413	769
P132	54	334	229	21	638
P212	227	0	273	218	718
P213	82	287	271	95	735
P221	74	281	74	339	768
P222	119	21	117	446	703
P322	68	349	2	185	604
P411	44	7	57	458	566
P423	7	14	39	576	636
P425	75	164	49	402	690
P431	129	0	0	530	659

In preprocessing, all words are stored into the term list. Dependency relations among words are considered. Each phrase is able to be considered as a set of words. A phrase P_i represents a vector as follows;

$$P_i = (w_1^{P_i}, \ldots, w_m^{P_i} \ldots, w_{N_{P_i}}^{P_i}), \quad (1)$$

where i is the index of each phrase, $w_m^{P_i}$ is the m-th word appeared in the phrase P_i, and N_{P_i} is the number of the words in the phrase P_i.

When a phrase P_i has a dependency relation to P_j, the dependency relation is able to be represented by a relation

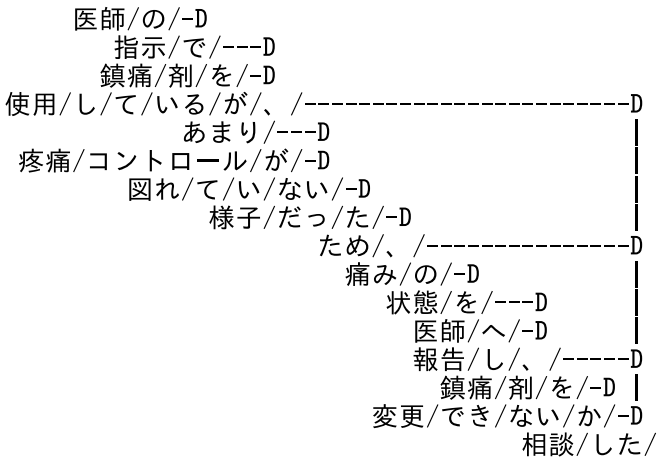

Fig. 1. Dependency relations extracted by Cabocha.

between the last element of P_i and the first element of P_j,

$$P_i \longrightarrow P_j \stackrel{def}{=} w^{P_i}_{N_{P_i}} \longrightarrow w^{P_j}_1, \tag{2}$$

where $\longrightarrow$ means a dependency relation. We can also consider every word in the same phrase has the dependency relation. The dependency relation between P_i and P_j can be represented as a word chain. The feature value corresponding to the word $w^{P_i}_m$ is the index of the next element $w^{P_i}_{m+1}$ and the value corresponding to the last element of a phrase P_i is the index of the first element of the next phrase P_j. Every dependency relation is encoded by using the index of the destination word.

The dependency relation based feature definition is worked well, if we know all words that appear in the test texts previously. However, there are several words that appear only in the test texts. In such case, since the source and/or destination words are missing, such missing words cause lack of features or missing feature values. To avoid the destination word missing issue, we use the existence information of the source word instead of the index of the destination word. In order to merge both the dependency relation and the existence information, the dependency relation vector is normalized into $[0,\ 1]$. Then the normalized dependency relation vector is added to a binary vector which represents word existence information. That is, the feature vector $\mathbf{x}_p$ takes the value of interval $[0,\ 2]$.

$$\mathbf{x}_p = (v_1, \ldots, v_k, \ldots, v_{N_{TL^{Q_l}_{year}}}), \tag{3}$$

$$v_k = \begin{cases} v, & \text{if the term which is corresponding to } v_k \\ & \text{exists and has the dependency relation,} \\ & 1 < v \leq 2, \\ 1, & \text{if the term which is corresponding to } v_k \\ & \text{exists and has no dependency relation,} \\ 0, & \text{if the term which is corresponding to } v_k \\ & \text{does not exist,} \end{cases}$$

where $N_{TL^{Q_l}_{year}}$ is the number of words in the term list $TL^{Q_l}_{year}$.

IV. K-NEAREST NEIGHBOR BASED SOFT CLASS DECISION

First, we assume that texts whose appearance of written words is similar should be classified into the same class. If the above assumption was true, it is enough to find the most similar text for classifying an evaluating text. However, the texts classified by experts, whose appearance of written words is similar, have sometimes different classes. In such case, we consider that the classification system can provide some candidate classes with certainty grade to human users.

K-nearest neighbor algorithms with fuzzy have been proposed such as in [13], [14]. These methods considered that the distance between vectors is used as weight for vector's membership, or is fuzzified for voting scheme.

In this paper, we consider soft class decision in classifying texts. The distance between vectors is used for deciding the number of nearest neighbors. The proportion of a class to the k nearest neighbors is considered as a certainty grade of that class.

In order to measure similarity of two feature vectors, we adopt cosine similarity.

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^N x_i y_i}{\sqrt{\sum_{i=1}^N x_i^2}, \sqrt{\sum_{i=1}^N y_i^2}}, \tag{4}$$
$$\boldsymbol{x} = (x_1, x_2, \ldots, x_N),$$
$$\boldsymbol{y} = (y_1, y_2, \ldots, y_N).$$

Let $\boldsymbol{x}^{\text{test}}_q$ be the feature vector of the evaluating text. Also let $\boldsymbol{x}^{\text{training}}_p$ be the feature vector of the p-th text in the training text set, θ be the threshold value for the cosine similarity, c^{training}_p be the class label of the p-th training text. The following is the proposed algorithm;

1) First, calculate cosine similarity values between x^{test}_q and x^{training}_p.

$$S_p = \cos(\boldsymbol{x}^{\text{training}}_p, \boldsymbol{x}^{\text{test}}_q), \tag{5}$$
$$p = 1, 2, \ldots, N_{\text{training}},\ q = 1, 2, \ldots, N_{\text{test}},$$
$$S = \{S_1, S_2, \ldots, S_{N_{\text{training}}}\}. \tag{6}$$

2) Find subset S'.

$$S' = \{S_p \mid S_p \geq \theta\}. \tag{7}$$

If $S' = \phi$, then reject x^{test}_q classification by this algorithm and classify it by using SVM based classification system proposed in [10].

3) Vote to every class.

$$v_c = \sum_{p \in S',\ c^{\text{training}}_p = c} w_p, \tag{8}$$
$$c = 1, 2, \ldots, N_{\text{class}},$$

where we define $w_p = 1$ in this paper.

4) Calcurate the certainty grade for each class CF_c.

$$CF_c = \frac{v_c}{\max(v_k | k = 1, 2, \ldots, N_{\text{class}})}. \tag{9}$$

TABLE II

CLASSIFICATION RESULTS BY SVM BASED SYSTEM

Q_l	# of texts	NCT*
P123	827	496 (60%)
P131	769	519 (67%)
P132	638	391 (61%)
P212	718	358 (50%)
P213	735	405 (55%)
P221	768	401 (52%)
P222	703	540 (77%)
P322	604	408 (68%)
P411	566	491 (87%)
P423	636	528 (83%)
P425	690	426 (62%)
P431	659	530 (80%)

NCT: the Number of Correctly classified Texts

5) Decide class for x_q^{test}. Class c_q for x_q^{test} is represented as follows.

$$x_q^{\text{test}} \text{ is classified into Class } c_q \text{ with } CF = CF_c. \tag{10}$$

If multiple classes have $CF > 0$, then such classes are presented as the class candidates.

In our proposed algorithm, the number of nearest neighbors k is determined by the similarity threshold θ. Since the number of nearest neighbors depends on the threshold θ, that number is varied for each test text x_q^{test}.

Our proposed method presents some class candidates with certainty grades. Human users may be confused when the number of candidates is large. We need to choose appropriate number of candidates for presenting classification results to human users.

V. EXPERIMENTAL RESULTS

In this section, some results were presented. Nursing-care text sets described in the section II were used as training and testing sets; (1) 2007 and 2008 text sets for training, and (2) 2009 text set for testing. First, both (1) and (2) text sets were converted into dependency based feature vectors by using the converting method described in the section III. Then, the proposed k-nearest neighbor based classification was applied to classify the test text set. The SVM based classification system for nursing-care text classification proposed in [1]–[10] was also used to classify rejected texts by k-nearest neighbor based system.

Table II shows a classification result when the SVM based classification system with dependency relation based feature vector definition was used. Q_l is the name of the text set. From this table, about 50–87% of texts were correctly classified by SVM based system.

Figures 2–9 show classification results for several values of threshold θ. For P222, P411, P425, and P431, the classification rate was not improved. When θ is lower, the number of nearest neighbors is larger. In this case, k-nearest neighbor

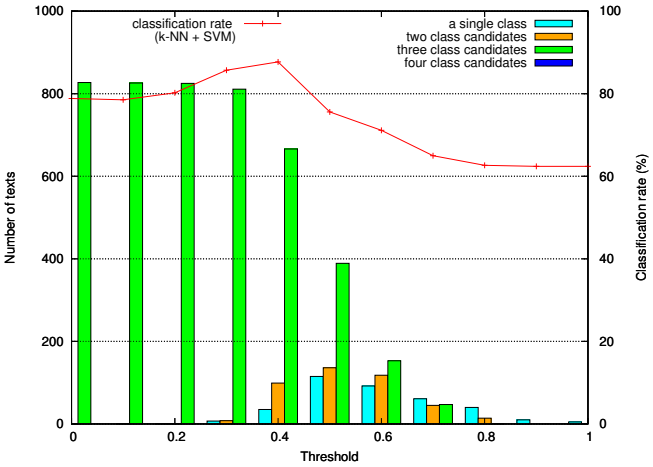

Fig. 2. Classification results for P123

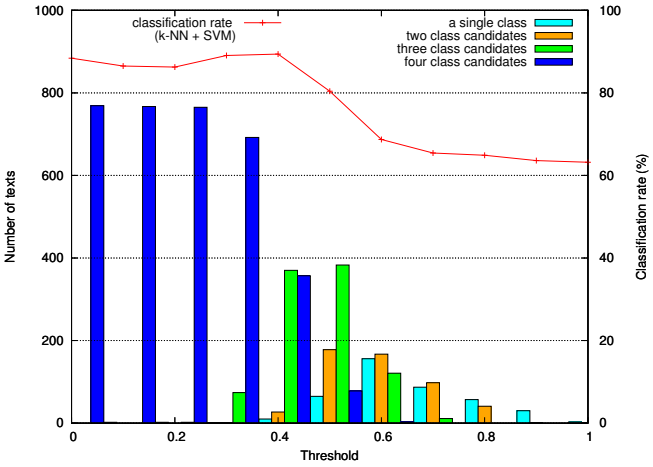

Fig. 3. Classification results for P131

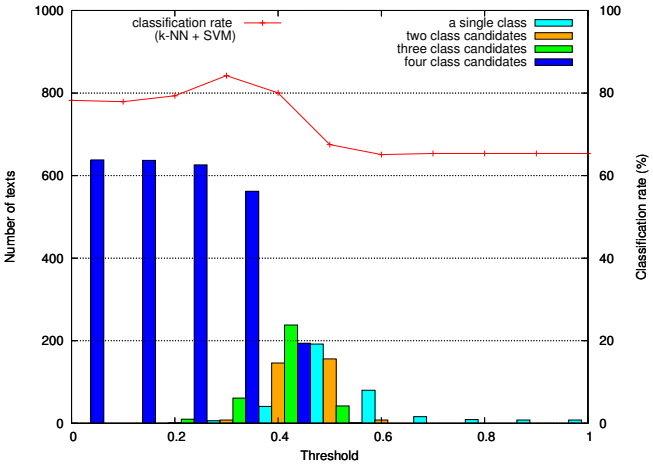

Fig. 4. Classification results for P132

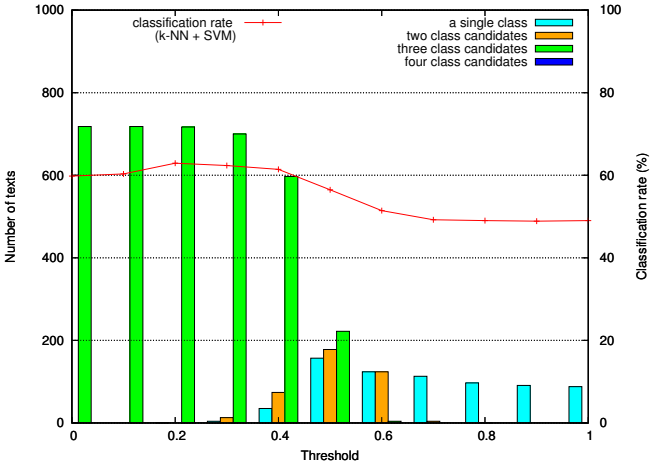

Fig. 5. Classification results for P212

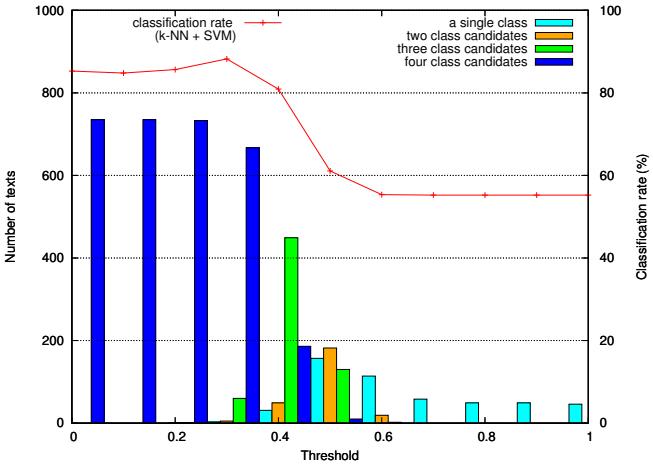

Fig. 6. Classification results for P213

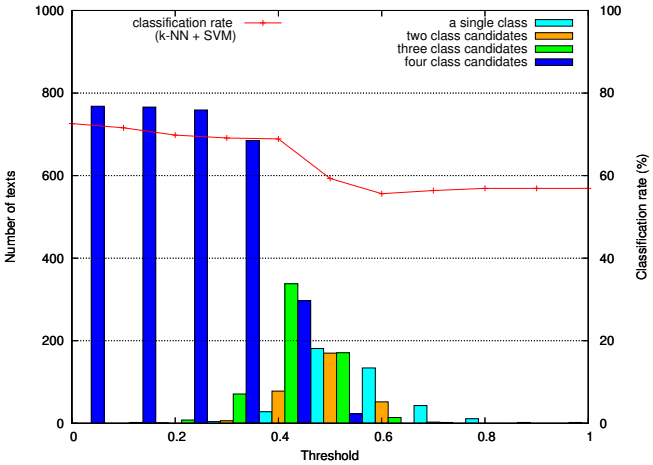

Fig. 7. Classification results for P221

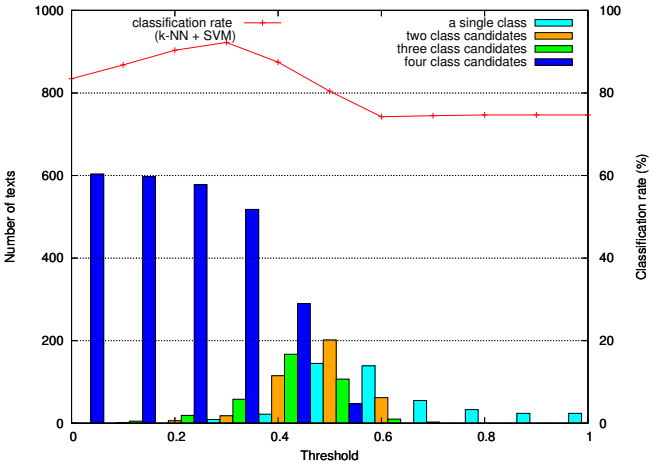

Fig. 8. Classification results for P322

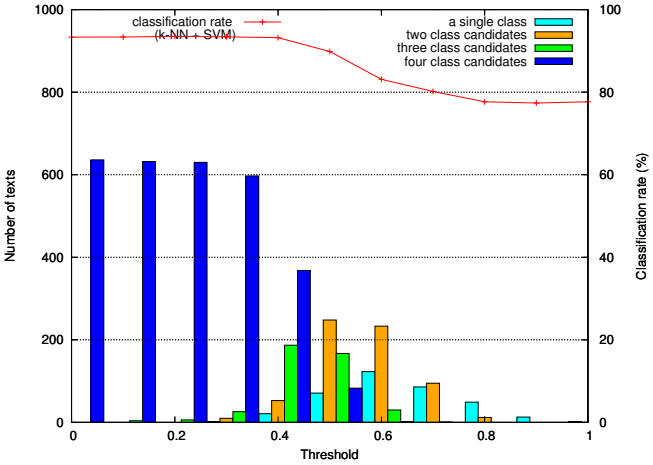

Fig. 9. Classification results for P423

based system presented some class candidates because the probability of including some classes is increasing. Although multiple class candidates increase the classification performance apparently, it does not make sense because almost texts have three or four classes. It is better that a few texts have class candidates and the total classification rate is higher. Therefore, about $\theta = 0.5$ is better than other θs for our purpose. Several hundreds of candidates are presented to experts if $\theta = 0.4, 0.5$ or 0.6 is chosen. When all texts are classified by experts, over five hundreds of texts have to be reviewed. Using our proposed system, experts need to review fewer texts than the above mentioned case. Also, since those texts have class candidates with certainty grades, experts can decide its class referring to the certainty grade.

VI. Conclusions

In this paper, we proposed a k-nearest neighbor based classification system which can classify into classes with certainty grade. To classify the nursing-care texts with high accuracy, a problem of handling such ambiguous class labels in the training data was tackled. In the proposed k-nearest

neighbor based system, nearest neighbors were selected by considering similarity threshold θ instead of k. Certainty grades of class candidates were decided by class distributions in the set of nearest neighbors. By choosing better balance between classification performance and the number of class candidates, we can develop a computer aided evaluation system to improve the quality of nursing-care.

One of our future works is to compare the proposed method with the other methods such as [13], [14]. Also we should examine the validity of class candidates represented by our proposed method.

REFERENCES

[1] M. Nii, S. Ando, Y. Takahashi, A. Uchinuno, and R. Sakashita, "Nursing-care Freestyle Texts Classification using Support Vector Machines", Proc. of 2007 IEEE International Conference on Granular Computing, CA, pp. 665–668.

[2] M. Nii, S. Ando, Y. Takahashi, A. Uchinuno, and R. Sakashita, "Feature extraction from nursing-care texts for classification", Proc. of 6th International Forum on Multimedia and Image Processing, 2008, in CDROM (6pages).

[3] M. Nii, S. Ando, Y. Takahashi, A. Uchinuno, and R. Sakashita, "GA based Feature Selection for Nursing-Care Freestyle Text Classification", Proc. of Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on advanced Intelligent Systems, 2008, pp. 756–761.

[4] M. Nii, T. Yamaguchi, Y. Takahashi, R. Sakashita and A. Uchinuno, "Analysis of Nursing-care Freestyle Japanese Text Classification Using GA-based Term Selection", Proc. of World Automation Congress 2010, Kobe, Japan, IFMIP495 (CD-ROM, 6 pages).

[5] M. Nii, T. Yamaguchi, Y. Takahashi, A. Uchinuno, R. Sakashita, "Improving Classification Performance of Nursing-Care Text Classification System by Using GA-based Term Selection", Journal of Advanced Computational Intelligence and Intelligent Information, Vol.14, No.2, pp. 142–149.

[6] M. Nii, T. Yamaguchi, Y. Mori, Y. Takahashi, R. Sakashita, and A. Uchinuno, "Classification of Nursing-care Data using Additional Term Information", Proc. of SCIS & ISIS 2010, Okayama, Japan, pp. 1469–1474.

[7] M. Nii, Y. Takahashi, A. Uchinuno, and R. Sakashita, "An Approach using Conceptual Fuzzy Sets for Nursing-care Text Classification", Proc. of 2012 World Automation Congress, Puerto Vallarta, Mexico, 2012, WAC 2012 1569535307.

[8] M. Nii, Y. Hirohata, A. Uchinuno, R. Sakashita, "New Feature Definition for Improvement of Nursing-care Text Classification", 2012 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2610–2615.

[9] M. Nii, Y. Hirohata, A. Uchinuno, R. Sakashita, "Feature Definition using Dependency Relations between Terms for Improving Nursing-care Text Classification", Proc. of 2012 Fifth International Conference on Emerging Trends in Engineering & Technology, pp. 110–115.

[10] M. Nii, S. Miyake, K. Takahama, A. Uchinuno, R. Sakashita, "Consideration about Utilizing Text Architecture for Making Feature Vectors in Classifying Nursing-care Texts", 2013 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1817–1821.

[11] MeCab, Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.jp/.

[12] Taku Kudo, Yuji Matsumoto, "Japanese Dependency Analysis using Cascaded Chunking", CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops), pp. 63–69.

[13] J. M. Keller, M. R. Gray, J. A. Givens, "A fuzzy k-nearest neighbor algorithm", IEEE Trans. Systems Man Cybernetics, vol. SMC-15, no. 4, pp. 580–585.

[14] H. B. Mitchell, P. A. Schaefer, "A "soft" K-nearest neighbor voting scheme", Int. J. Intell. Syst., 16: 459468. doi: 10.1002/int.1018