Marketta Hiissa | Tapio Pahikkala | Hanna Suominen | Tuija Lehtikunnas | Barbro Back | Helena Karsten | Sanna Salanterä | Tapio Salakoski

Towards Automated Classification of Intensive Care Nursing Narratives

Turku Centre for Computer Science

TUCS Technical Report
No 739, January 2006

Towards Automated Classification of Intensive Care Nursing Narratives

Marketta Hiissa
Turku Centre for Computer Science and Åbo Akademi University
marketta.hiissa@abo.fi

Tapio Pahikkala
Turku Centre for Computer Science and University of Turku
tapio.pahikkala@utu.fi

Hanna Suominen
Turku Centre for Computer Science and University of Turku
hanna.suominen@utu.fi

Tuija Lehtikunnas
Turku University Hospital and University of Turku
tuija.lehtikunnas@utu.fi

Barbro Back
Turku Centre for Computer Science and Åbo Akademi University
barbro.back@abo.fi

Helena Karsten
Turku Centre for Computer Science and University of Turku
eija.karsten@utu.fi

Sanna Salanterä
University of Turku, Department of Nursing Science
sanna.salantera@utu.fi

Tapio Salakoski
Turku Centre for Computer Science and University of Turku
tapio.salakoski@utu.fi

TUCS Technical Report

No 739, January 2006

Abstract

Nursing narratives are an important part of patient documentation, but the possibilities to utilize them in the direct care process are limited due to the lack of proper tools. One solution to facilitate the utilization of narrative data could be to classify them according to their content. In this paper, we addressed two issues related to designing an automated classifier: the agreement on the content of the classes into which the data are to be classified, and the ability of the machine-learning algorithm to perform the classification on an acceptable level. The data we used were a set of Finnish intensive care nursing narratives. By using Cohen's κ, we assessed the agreement of three nurses on the content of the classes Breathing, Blood Circulation and Pain, and by using the area under ROC curve (AUC), we measured the ability of the Least Squares Support Vector Machine (LS-SVM) algorithm to learn the classification patterns of the nurses. On average, the values of κ were around 0.8. The agreement was highest in the class Blood Circulation, and lowest in the class Breathing. The LS-SVM algorithm was able to learn the classification patterns of the three nurses on an acceptable level; the values of AUC were generally around 0.85. Our results indicate that one way to develop electronic patient records could be tools that handle the free text in nursing documentation.

Keywords: Computerized Patient Records, Natural Language Processing, Nursing

TUCS Laboratory
Data Mining and Knowledge Management Laboratory
Bioinformatics Laboratory
Health and Medical Informatics Institute

1 Introduction

During the past years, health-care providers have been changing paper-based patient records to electronic ones. This has, on one hand, made more data available on each patient, but on the other hand, also offered new possibilities to utilize the gathered data. However, the effects of this switch have not only been positive. It has been found that electronic charting may not provide nurses with more time for tasks unrelated to manipulating data [1, 2], and that electronic systems support nurses in gathering information, but not in the active utilization of it [3].

Especially problematic is the active utilization of narrative data, in particular when the patient stays in the ward for several days, and the amount of documentation is large. In intensive care units, a variety of artificial intelligence techniques can already be used to process the numerical or structured data in the patient records (e.g. [4]), but due to the lack of proper tools, the possibilities to utilize narrative data are still limited. Our approach to facilitate the utilization of narratives is to develop automatic tools that classify texts, and thus make it easier to retrieve relevant information e.g. when a nurse needs to build a general picture of narratives describing patient's breathing.

In the medical domain, classification has recently been applied e.g. to classifying texts such as chief complaint notes, diagnostic statements, and injury narratives into different kinds of syndromic, illness and cause-of-injury categories [5, 6, 7, 8]. There is, however, little research on the automated processing of nursing narratives [9].

In this paper, we use a machine-learning approach, i.e. an algorithm that learns the classification patterns directly from pre-classified data, to classify Finnish intensive care nursing narratives. We address two issues related to designing a classifier: the agreement on the content of the classes into which the data are to be classified, and the ability of the classification algorithm to perform on an acceptable level. The classes used in this study are Breathing, Blood Circulation and Pain, and the machine-learning algorithm is the Least Squares Support Vector Machine algorithm.

2 Material and Methods

2.1 Material

The data we used were a set of Finnish intensive care nursing narratives. The documents were gathered, with proper permissions and without any identification information, in the spring of 2001 from 16 intensive care units, two or three documents per unit. In total, we had 43 copies of anonymous patient records with nursing notes written down during one day and night.

The style of the documentation varied from one nurse to another: some had written short sentences such as ”*Hemodynamics ok. Very tired.*”, whereas others had written long sentences in which different matters were separated with commas. In order to standardise the style of the documentation, we divided the long sentences into smaller pieces consisting of one matter or thought. This was done manually by one of the authors with nursing experience, and resulted in 1363 pieces, with the average length of 3.7 words.

2.2 Methods

We chose to classify the data according to classes Breathing, Blood Circulation and Pain. The selection of breathing and blood circulation was due to the emphasis in intensive care nursing, which is on the monitoring, assessment and maintenance of vital functions [10]. Pain was chosen because due to their critical illnesses, patients are often incapable of communicating, and nurses must assess the level of pain relying only on different kinds of behavioural and physiological indicators [11], and thus the documentation of pain was supposed to differ from the one of breathing or blood circulation. The classification process was done as three separate classification tasks, i.e. each of the classes was considered separately of the others.

In order to assess the agreement on the content of the three classes, we asked three nurses to manually classify the data, independently of each other. They were advised to label each text piece they considered to contain useful information given the specific class. All the nurses were specialists in nursing documentation; two had a long clinical experience from intensive care units (N_1 and N_2), and one was an academic nursing science researcher (N_3). We measured the agreement on the content of the classes by using Cohen's κ [12]. Cohen's κ is a chance-corrected measure of agreement, which considers the classifiers equally competent to make judgments, places no restriction on the distribution of judgments over classes for each classifier, and takes into account that a certain amount of agreement is to be expected by chance. It is an appropriate measure especially in situations like this when there are no criteria for correct classification.

In order to test the performance of the machine-learning algorithm, we divided the data classified by the nurses into a training set and a test set so that 708 out of the total of 1363 text pieces belonged in the training set, and the remaining 655 pieces in the test set. The division was done so that text pieces from one document belonged only to one of the two sets. The machine-learning algorithm we used was a Least Squares Support Vector Machine (LS-SVM) [13], which is a technique that has been shown to yield good results on classification problems (e.g. [14]). Nine automated classifiers were trained by using the training data labelled by the three nurses, i.e. one classifier for each nurse-class pair. In order to reduce different inflection forms of the words, the data were pre-processed with the Snowball stemmer for Finnish [15]. The performance of the algorithm was measured with respect to the test data by using the area under ROC curve (AUC) [16], which corresponds to the probability that given a randomly chosen positive example and a randomly chosen negative example, the classifier will correctly determine which is which.

All the statistical analyses were performed with SPSS 11.0 for Windows.

3 Results

3.1 Agreement on the content of the classes

The amount of data the nurses included in the classes Breathing, Blood Circulation and Pain was, respectively, around 20%, 15%, and 6% of the 1363 text pieces. This reflects the extensive content of the narrative documentation, and illustrates the difficulty of finding relevant information from large amounts of text.

Table 1: The values of Cohen's κ and the respective 95% confidence intervals (CI) for the agreement between the three nurses.

	Breathing	Blood Circulation	Pain
	κ 95% CI	κ 95% CI	κ 95% CI
$N_1 - N_2$	0.73 (0.68-0.78)	0.89 (0.85-0.92)	0.88 (0.82-0.94)
$N_1 - N_3$	0.67 (0.62-0.72)	0.81 (0.77-0.86)	0.79 (0.73-0.86)
$N_2 - N_3$	0.85 (0.82-0.89)	0.87 (0.83-0.90)	0.76 (0.69-0.83)

The comparisons between the nurses showed that the text pieces describing blood circulation were selected quite similarly ($\kappa > 0.80$ in each comparison), whereas there were more differences in selecting text pieces related to pain or breathing (Table 1). The range of the values of κ was the largest, from 0.67 to 0.85, in the class Breathing. In addition, given the classes Pain and Blood Circulation, the two clinical nurses N_1 and N_2 where the most unanimous in the classification, whereas in the class Breathing, the clinical nurse N_2 and the nursing science researcher N_3 were the most unanimous.

3.2 Learning ability of the LS-SVM algorithm with respect to the data classified by the nurses

The learning ability of the LS-SVM was tested in two ways. Firstly, we tested the LS-SVM classifiers against test data classified by the same nurse whose data was used when training the algorithm. The results showed that the algorithm was able to learn the classification patterns from the training set (Table 2). The values of AUC were in general around 0.85. The highest values, from 0.89 to 0.93, were achieved for the classification of blood circulation statements, whereas the lowest, from 0.71 to 0.81 were measured for the class Pain. Except the class Pain, the performance of the algorithm was on a similar level in the classes independently of the nurse whose classification was used to train the algorithm, also in the class Breathing, in which the disagreement between the nurses was the highest.

Table 2: The values of AUC and the respective 95% confidence intervals (CI) for the automatic classifiers.

	Breathing		Blood Circulation		Pain	
	AUC	95% CI	AUC	95% CI	AUC	95% CI
N_1	0.86	(0.82-0.90)	0.89	(0.84-0.93)	0.71	(0.61-0.80)
N_2	0.88	(0.85-0.91)	0.93	(0.90-0.97)	0.81	(0.73-0.89)
N_3	0.87	(0.84-0.91)	0.91	(0.86-0.95)	0.71	(0.61-0.80)

Secondly, we tested the classifiers against test data classified by other nurses than the one whose data was used when training the algorithm. The values of AUC were calculated for the total of six comparisons in each of the three classes, and based on these comparisons, we measured the average decrease in the values of AUC compared with the situation in which both the training and the testing data were classified by the same nurse. The average decrease in the values was 0.06 in the class Breathing, and 0.01 in the class Pain. In the class Blood Circulation, the decrease was 0.00, i.e. given a manual classification, the two LS-SVM classifiers trained with other data than that of the given nurse performed as well as the classifier trained by using the classification of the given nurse.

4 Discussion

We have here assessed the agreement of three nurses on the content of the classes Breathing, Blood Circulation and Pain by using Cohen's κ, and the ability of the LS-SVM machine-learning algorithm to learn the classification patterns of the nurses by using AUC as the outcome measure. On average, the values of κ were around 0.8. According to the obtained AUC values, the LS-SVM algorithm performed on an acceptable level; the values were generally around 0.85, and the decreases in these were rather small when using test data classified by another nurse than the one who classified the training data.

The disagreement cases between the nurses appeared to be due to not only the subjective considerations and interpretations on what information was important given a specific class, but also to matters such as the different handling of non-standard abbreviations. For example, one of the nurses decided not to accept any text pieces containing non-standard and ambiguous abbreviations to belong in any class, whereas the others classified these pieces based on the meaning they thought the abbreviations could have in the context they appeared. The effects of the subjective considerations on important information were the most visible in the classes Breathing and Pain. For example, given the class Breathing, nurses disagreed on whether or not the statements related to the slime in patient's lungs should be included in the class, and given the class Pain, the disagreements were mainly due to sentences such as ”*Reacts to interventions by moving extremities and furrowing eyebrows*.”, which did not include any direct mention of pain.

According to the obtained values of AUC, the learning ability of the LS-SVM algorithm was on a rather high level. The results for the classes Blood Circulation and Breathing were good, and the performance in the class Pain was somewhat lower mainly due to the small amount of available data and the different nature of the documentation. The sentences in the classes Blood Circulation and Breathing contained many keywords, such as heart rate, blood pressure, and hemodynamics in the class Blood Circulation, and oxygen saturation, respirator, and intubation in the class Breathing. In contrast to these, the class Pain included more non-direct statements, in which there were no keywords present. Compared to the other nurses, the nurse N_2 included fewer non-direct statements in the class Pain, which explains the differences between the automated classifiers in that class. The results also showed that in the class Blood Circulation, there were no decreases in the values of AUC when using test data classified by another nurse than the one with whose classification the algorithm was trained. The decreases in the values of AUC in the classes Breathing and Pain reflect the nurses' agreement on the content of these classes; in the class Blood Circulation, the values of where higher than in the classes Breathing and Pain.

In order for the machine-learning based classification to be useful in the intended application area, the performance of the algorithm needs to be on a satisfactory level, but also the training data needs to be well established. In this establishment, the disagreement cases between the nurses can be handled in different ways. On one hand, they could be taken into further consideration in order to reach a consensus on them, but on the other hand, if they are considered to include valuable information about the domain, the utilization of them in establishing the training data is justified. For example, our results revealed that with the pain statements, all the nurses did not include in the class expressions such as ”*Reacts against when turning him over*.”, in which the nurse does not explicitly denote the presence of pain. These disagreement cases could be considered e.g. as weaker pain statements than those on which all the nurses agreed, and this information could be used to give different weights to different kinds of statements with respect to the given class.

Our results showed that the LS-SVM algorithm performed on an acceptable level. However, improvements could be gained e.g. by using more training data, and by increasing the pre-processing of the data. Here we used a stemmer to reduce different inflection forms of the words, but because Finnish is a highly inflectional language, techniques that find the real base form instead of just stemming could make the data less sparse and increase the performance of the algorithm. Another topic of further research is the automation of the pre-processing of the long sentences, which here was done manually by one researcher, and is thus a possible source of bias. Further research is also needed to assess the performance of the algorithm on other classes than the three used in this study.

5 Conclusion

Given the classes Breathing, Blood Circulation and Pain, the nurses had somewhat different opinions on the optimal content of them, and these differences could be utilized when designing an automated classifier. The LS-SVM algorithm was able to learn classification patterns from the data classified by the nurses, and its performance on unseen material was on an acceptable level. We conclude that one way to develop electronic patient records could be tools that handle the free text in nursing documentation.

Acknowledgments

We gratefully acknowledge the financial support of Tekes (grant number 40435/05) and the Academy of Finland (grant number 104639). We wish to thank Heljä Lundgren-Laine and Filip Ginter for their contributions.

References

[1] Pierpont GL, Thilgen D. Effect of computerized charting on nursing activity in intensive care. *Critical Care Medicine* 1995; 23(6): 1067–1073.

[2] Smith K, Smith V, Krugman M, Oman K. Evaluating the impact of computerized clinical documentation. *Computers, Informatics, Nursing* 2005; 23(3): 132-138.

[3] Snyder-Halpern R, Corcoran-Perry S, Narayan S. Developing clinical practice environments supporting the knowledge work of nurses. *Computers in Nursing* 2001; 19(1): 17-23; quiz 24-6.

[4] Hanson CW,3rd, Marshall BE. Artificial intelligence applications in the intensive care unit. *Critical Care Medicine* 2001; 29(2): 427-435.

[5] Chapman WW, Dowling JN, Wagner MM. Fever detection from free-text clinical records for biosurveillance. *Journal of Biomedical Informatics* 2004; 37(2): 120-127.

[6] Pakhomov SV, Buntrock JD, Chute CG. Using compound codes for automatic classification of clinical diagnoses. *Medinfo* 2004; 11(Pt 1): 411-415.

[7] Wellman HM, Lehto MR, Sorock GS, Smith GS. Computerized coding of injury narrative data from the National Health Interview Survey. *Accident; Analysis and Prevention* 2004; 36(2): 165-171.

[8] Chapman WW, Christensen LM, Wagner MM, Haug PJ, Ivanov O, Dowling JN, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. *Artificial Intelligence in Medicine* 2005; 33(1): 31-40.

[9] Bakken S, Hyun S, Friedman C, Johnson SB. ISO reference terminology models for nursing: applicability for natural language processing of nursing narratives. *International Journal of Medical Informatics* 2005; 74(7-8): 615-622.

[10] Ward NS, Snyder JE, Ross S, Haze D, Levy MM. Comparison of a commercially available clinical information system with other methods of measuring critical care outcomes data. *Journal of Critical Care* 2004; 19(1): 10-15.

[11] Payen JF, Bru O, Bosson JL, Lagrasta A, Novel E, Deschaux I, et al. Assessing pain in critically ill sedated patients by using a behavioral pain scale. *Critical Care Medicine* 2001; 29(12): 2258-2263.

[12] Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; 20(1): 37-46.

[13] Suykens JAK, Van Gestel T, De Brabanter, J, De Moor B, Vandewalle J. *Least Squares Support Vector Machines*. Singapore: World Scientific Publishing, 2002.

[14] R. M. Rifkin. *Everything Old is New Again: A Fresh Look at Historical Approaches in Machine Learning*. PhD Thesis, Massachusetts Institute of Technology; 2002.

[15] http://snowball.tartarus.org/algorithms/finnish/stemmer.html (last accessed Jan 10, 2006).

[16] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143(1): 29-36.

Turku

Centre *for*

Computer

Science

Lemminkäisenkatu 14 A, 20520 Turku, Finland | www.tucs.fi

University of Turku

- Department of Information Technology
- Department of Mathematics

Åbo Akademi University

- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration

- Institute of Information Systems Sciences

ISBN 952-12-1682-4
ISSN 1239-1891