Driving Reform: Digital Health is Everyone's Business
A. Georgiou et al. (Eds.)

doi:10.3233/978-1-61499-558-6-87

Automated Classification of Clinical Incident Types

Jaiprakash GUPTA [1], Irena KOPRINSKA and Jon PATRICK
School of Information Technologies, University of Sydney, Australia

Abstract. We consider the task of automatic classification of clinical incident reports using machine learning methods. Our data consists of 5448 clinical incident reports collected from the Incident Information Management System used by 7 hospitals in the state of New South Wales in Australia. We evaluate the performance of four classification algorithms: decision tree, naïve Bayes, multinomial naïve Bayes and support vector machine. We initially consider 13 classes (incident types) that were then reduced to 12, and show that it is possible to build accurate classifiers. The most accurate classifier was the multinomial naïve Bayes achieving accuracy of 80.44% and AUC of 0.91. We also investigate the effect of class labelling by an ordinary clinician and an expert, and show that when the data is labelled by an expert the classification performance of all classifiers improves. We found that again the best classifier was multinomial naïve Bayes achieving accuracy of 81.32% and AUC of 0.97. Our results show that some classes in the Incident Information Management System such as Primary Care are not distinct and their removal can improve performance; some other classes such as Aggression Victim are easier to classify than others such as Behavior and Human Performance. In summary, we show that the classification performance can be improved by expert class labelling of the training data, removing classes that are not well defined and selecting appropriate machine learning classifiers.

Keywords. Clinical incidents, patient safety, machine learning, statistical text classification, decision tree, naïve Bayes, naïve Bayes multinomial, support vector machine

Introduction

In this paper we consider the task of automatic classification of clinical incident reports. Manual coding of free text documents is expensive and inefficient, especially when applied to big datasets. An automated solution that is accurate and consistent is highly desirable [1]. Correct encoding and reporting of incident types is required for patient safety and health system improvement [2, 3]. The architecture of the Incident Information Management System (IIMS) used in Australia is very complex but its core, the Generic Reference Model, has never been tested with statistical rigor [4, 5]. In the state of New South Wales there are over a million clinical incidents documented in IIMS [5]. Our review and those done by others [6] found that research in the area of automated classification of is very contextual. There has been very little research on using statistical text classifiers on IIMS datasets, with only three reported studies [4, 6, and 7]. Ong et al. [7] used binary models for investigating clinical handover, patient identification and risk category instances [8]. In this paper, for the first time, we

[1] Corresponding Author: Jai Gupta; Email: jgup5981@uni.sydney.edu.au.

consider classification of incident reports into multiple classes; in particular, we consider 12-13 clinical incident types. Another contribution of our work is evaluating the performance of more machine learning classifiers than in previous work [7], to determine the best classifier, using a comprehensive set of accuracy measures. We also study the effect of document labelling undertaken by a clinician compared to an expert, on the accuracy of the classification.

1. Method

We used IIMS data collected from January 2004 to December 2008 that contains information about the following 13 Clinical Incident Types (CIT): Aggression Aggressor (AA), Aggression Victim (AV), Blood and Blood Product (BBP), Behavior and Human Performance (BHP), Clinical Management (CM), Documentation (DOC), Fall (FALL), Hospital Associated Infection/infestation (HAI), Medication (MED), Nutrition (NUT), Primary Care (PM), Nutrition (NUT), Primary Care (PC), Pathology Lab (PATH) and Pressure Ulcers (PU). The total number of incidents was 5448, 250 for each category except for categories HAI, NUT, PATH and PC, where the number of documents was 361, 250, 306 and 31, respectively.

We applied four machine learning classifiers available from WEKA [9], an open source data mining software: Decision Trees (DT), Naïve Bayes (NB), Naïve Bayes Multinomial (NBM) and Support Vector Machine with radial basis kernel function (SVM_RBF). We chose them as they represent different machine learning paradigms and are also state-of-the-art classifiers. We compared their performance by computing the following standard accuracy measures: overall accuracy, recall, precision, F1 measure, Area Under the Curve (AUC) and Kappa statistic.

We conducted four experiments. Experiment 1 used 13 CIT, 14 fields of information and 5448 reports; Experiment 2 used 12 CIT (incident type PC was removed), 10 fields and 5417 reports and Experiment 3 used 12 CIT, 10 fields and 1200 clinical incidents (100 reports per CIT). In all these three experiments the incident reports were classified by clinicians. Experiment 4 was the same as Experiment 3 but the incident reports were classified by an expert not a clinician (the first author of the paper who has over 10 years of experience using IIMS). We followed the same experimental methodology as in [5], including the methods for feature extraction, selection and representation.

2. Results

2.1. Experiments 1 and 2

Table 1 presents the accuracy results for Experiment 1 (columns “13 classes”). The two best performing classifiers were SVM_RBF and NBM achieving accuracy of 78.29-79.06% and Kappa statistic of 0.76-0.77.

An examination of the confusion matrix (Table 2) for the least accurately predicted class PC revealed that it is often misclassified as one of the other classes. As the number of instances in this class is small (31), we removed it and conducted the evaluation using 12 classes (see Table 1, columns “12 classes”). This resulted in improved classification performance for all classifiers except for SVM_RBF – its

accuracy declined from 79.06% to 68.89%, in spite of maintaining the same precision (0.79). The most accurate classifier was NBM achieving accuracy of 80.44%, an improvement from 78.29%.

Table 1. Accuracy results of the four classifiers in Experiment 1 (13 CIT, N=5448) and Experiment 2 (12 CIT, N=5417), clinician classified CITs.

Algorithms	**DT**		**NB**		**NBM**		**SVM_RBF**	
CIT	13	12	13	12	13	12	13	12
Accuracy [%]	73.66	75.54	69.71	71.86	78.29	80.44	79.06	68.89
Kappa statistic	0.71	0.73	0.67	0.69	0.76	0.79	0.77	0.66
Precision	0.74	0.74	0.71	0.71	0.79	0.72	0.79	0.79
AUC	0.89	0.89	0.90	0.90	0.96	0.91	0.89	0.89

Table 2. Confusion matrix for class PC. The correctly classified instances are in bold, the rest are misclassifications.

Classified as – Class >		**AA**	**AV**	**BBP**	**BHP**	**CM**	**DOC**	**FALL**	**HAI**	**MED**	**NUT**	**PC**	**PATH**	**PU**
PC	DT	2	0	0	5	1	0	0	0	0	1	2	0	0
	NB	8	2	5	11	11	2	7	2	10	1	3	2	3
	NBM	0	1	1	4	2	1	2	0	2	0	3	0	0
	SVM_RBF	0	0	0	6	5	2	0	0	0	0	3	0	2

Table 3 shows the accuracy results for the most improved classes (BBP and AV) and the least improved class (BHP), when the number of classes was decreased from 13 to 12. The accuracy of the removed class (PC) is also shown for comparison. The highest improvement in recall was achieved by the DT classifier for class AV (from 0.79 to 0.93). The highest improvement in precision was achieved by the SVM_RBF classifier for class BBP (from 0.83 to 0.91). In terms of F1 measure, the highest improvement was achieved by the DT classifier for class AV (from 0.79 to 0.92). The highest improvement in terms of AUC was achieved by the DT classifier (from 0.95 to 0.98), and we note that 0.95 was already a very high accuracy.

2.2. Experiments 3 and 4

In these experiments a balanced set of clinical incident reports is used 100 reports per CIT, for the 12 CIT (N=1200 clinical incidents). In both experiments the 100 reports per class were randomly selected from the set of all available reports. In Experiment 3, the reports were classified by an ordinary clinician and in Experiment 4 they were classified by an expert. The random selection of 100 reports per class was conducted separately for the two experiments.

Table 4 presents the classification results. As we can see the expert classification resulted in an improvement for all classifiers, for all performance measures, except in 4

out of all 16 cases. In 2 of these cases the performance didn't change (Kappa statistic and precision for NB) and in the remaining 2 cases there was a decrease (AUC for NB and SVM_RBF). For example, in terms of accuracy the improvements were: 5% for DT, 1% for NB and NBM and 14% for SVM_RBF. The best results were achieved by NBM and the most improved classifier was SVM_RBF.

Table 3. Comparison between Experiment 1 (13 classes) and Experiment 2 (12 classes), results for the least and most improved classes.

Performance Measures		**Least Improved Classes**			**Most Improved Classes**				**Weighted Average**	
	Class	PC	BHP		BBP		AV			
		13	13	12	13	12	13	12	13	12
Recall	DT	0.07	0.47	0.45	0.80	0.80	0.79	0.93	0.74	0.76
	NB	0.10	0.46	0.48	0.73	0.73	0.74	0.77	0.70	0.72
	NBM	0.10	0.55	0.58	0.80	0.81	0.82	0.84	0.78	0.80
	SVM_RBF	0.10	0.59	0.68	0.84	0.61	0.86	0.80	0.79	0.69
Precision	DT	0.18	0.50	0.50	0.78	0.76	0.85	0.91	0.74	0.76
	NB	0.05	0.49	0.50	0.84	0.78	0.63	0.70	0.71	0.73
	NBM	0.19	0.62	0.64	0.91	0.91	0.67	0.71	0.79	0.81
	SVM_RBF	0.17	0.59	0.38	0.83	0.91	0.89	0.64	0.79	0.75
F measure	DT	0.71	0.66	0.47	0.49	0.78	0.79	0.92	0.74	0.76
	NB	0.06	0.48	0.49	0.78	0.75	0.68	0.73	0.70	0.72
	NBM	0.13	0.59	0.61	0.85	0.86	0.74	0.77	0.78	0.80
	SVM_RBF	0.12	0.59	0.49	0.84	0.73	0.87	0.71	0.79	0.69
AUC	DT	0.60	0.77	0.79	0.89	0.89	0.95	0.98	0.89	0.90
	NB	0.64	0.79	0.80	0.93	0.92	0.90	0.90	0.90	0.90
	NBM	0.80	0.91	0.91	0.97	0.97	0.96	0.96	0.96	0.97
	SVM_RBF	0.55	0.78	0.78	0.91	0.80	0.93	0.87	0.89	0.83

Table 5 shows the results for the most and least improved CITs. The two most improved classes were AV and AA, and the two least improved classes were BHP and BBP. The highest improvement in recall, precision and AUC was achieved by DT for class AV (from 0.33 to 0.87, from 0.38 to 0.81, and from 0.77 to 0.97, respectively). In terms of F1 measure, the highest improvement was achieved by NBM for class AV (from 0.36 to 0.69). For comparison, for the least improved class, BHP, the DT classifier achieved a much smaller improvement (from 0.02 for AUC to 0.11 for recall).

Table 4. Results of the four classifiers in Experiment 3 (12 CIT, N=1200, clinician classified CITs) and Experiment 4 (12 CIT, N=1200, expert classified CITs).

Algorithms	**DT**		**NB**		**NBM**		**SVM_RBF**	
	Expert	Clinician	Expert	Clinician	Expert	Clinician	Expert	Clinician
Accuracy [%]	70.17	65.91	70.08	69.60	81.32	79.58	54.92	41.12
Kappa statistic	0.68	0.63	0.67	0.67	0.80	0.78	0.51	0.33
Precision	0.70	0.66	0.71	0.71	0.81	0.8	0.69	0.63
AUC	0.89	0.85	0.89	0.91	0.97	0.96	0.41	0.66

Table 5. Comparison between Experiment 3 (clinician classified CITs) and Experiment 4 (expert classified CITs) for 12 classes (N=1200).

Performance Measures		**Minimally improved Classes**				**Most improved Classes**				**Weighted Average**	
	Classes	BHP		BBP		AV		AA			
	Classifier	Exp	Cli	Exp	Cli	Exp	Cli	Exp	Cli	Exp	Cli
Recall	DT	0.67	0.56	0.75	0.68	0.87	0.33	0.84	0.35	0.70	0.66
	NB	0.48	0.44	0.76	0.70	0.74	0.54	0.66	0.35	0.71	0.71
	NBM	0.62	0.57	0.89	0.89	0.79	0.40	0.58	0.75	0.81	0.80
	SVM_RBF	0.42	0.37	0.47	0.24	0.74	0.56	0.39	0.35	0.69	0.63
Precision	DT	0.49	0.42	0.75	0.66	0.81	0.38	0.52	0.62	0.70	0.66
	NB	0.49	0.45	0.84	0.76	0.60	0.26	0.67	0.60	0.70	0.70
	NBM	0.64	0.63	0.93	0.86	0.61	0.33	0.75	0.62	0.81	0.80
	SVM_RBF	0.52	0.48	0.96	0.90	0.58	0.42	0.64	0.21	0.55	0.41
F1 measure	DT	0.78	0.73	0.43	0.41	0.75	0.67	0.81	0.79	0.00	0.00
	NB	0.83	0.77	0.46	0.46	0.80	0.73	0.90	0.88	0.00	0.00
	NBM	0.93	0.91	0.99	0.87	0.69	0.36	0.96	0.93	0.79	0.81
	SVM_RBF	0.50	0.25	0.63	0.39	0.59	0.00	0.80	0.69	0.56	0.41
AUC	DT	0.91	0.89	0.91	0.85	0.97	0.77	0.90	0.80	0.89	0.85
	NB	0.90	0.89	0.94	0.93	0.90	0.86	0.89	0.82	0.91	0.90
	NBM	0.95	0.95	0.98	0.97	0.95	0.90	0.95	0.91	0.97	0.96
	SVM_RBF	0.68	0.47	0.62	0.47	0.52	0.50	0.60	0.60	0.61	0.46

3. Discussion

Automatic classification of clinical incident reports into two classes has been successfully tested and reported in [5]. In this paper we considered 12-13 classes and showed that it is possible to build accurate multiclass classifiers of clinical incident

reports using machine learning methods. This is an important step as complex systems such as IIMS cannot be over simplified and require multiclass classification methods.

In particular, our results showed that the NBM classifier (which hasn't been previously applied to IIMS data) was consistently the best performing classifier, obtaining accuracy of 80.44% on clinician-labelled data and 82.32% on expert-labelled data. The second best classifiers were DT and NB, and finally SVM_RBF.

Kappa statistic gives a chance-corrected measure of agreement between the classifications and the true classes, 1indicates a statistically perfect modelling whereas a 0 means every model value was different from the actual value. Kappa statistic was 0.70 and over on NBM classifier. A kappa statistic of 0.70 or higher is generally regarded as good and getting close to statistically perfect model [10] and this was reached when combined data types were used.

An advantage of multiclass classification over binary classification is that it provides more useful insights which classes are difficult and easy to classify, and this knowledge can be used to improve the definition of the classes, in collaboration with clinical experts. We were able to identify class PC as frequently misclassified and showed that its removal improved the classification results. Prompted by the results, we found that class PC is not well defined and not clearly distinguishable from the other classes. Therefore, we recommend that it is removed from the IIMS and that it is recorded in a separate database.

4. Conclusion

In this paper we considered the task of automatic classification of clinical incident reports collected from an incident information management system. We formulated the problem as multiclass text classification and evaluated the performance of several machine learning classifiers. We found that NBM was the most accurate classifier achieving an accuracy of 80.44% on clinician-labelled data 82.32% on expert-labelled data, which are promising results. Our results showed that some classes such as Primary Care are often misclassified as they are not well defined, and we recommend their removal from the incident management system. We also found that expert-labelled training data resulted in better classification performance compared to clinician-labelled data.

References

[1] AHIMA computer assisted coding e-HIM work group. Delving into computer-assisted coding. J AHIMA, 75 (2010), 48A-48H.

[2] D. Pettet and L. Donaldson, Challenging the world: patient safety and health care associated infection. International J Quality in Health Care 18 (2006), 4-6.

[3] K.E. Wood and D.B.Nash, Mandatory state-based error-reporting systems: current and future prospects. Am J Med Qual 20 (2005), 297–303.

[4] W.B. Runciman, S.C.Helps, E.J. Sexton et al. A classification for incidents and accidents in the health-care system. J Qual Clin Pract.18 (1998),199–211.

[5] J. Gupta and J.Patrick Automated validation of patient safety clinical incident classification: macro analysis. Stud Health Technol Inform.(2013)188:52-7.

[6] M.H. Stanfill, M.Williams,S.H. Fenton, R.A.Jenders and W.R.Hersh, A systematic literature review of automated clinical coding and classification systems, J Am Med Inform Assoc. 17 (2010), 646-651. DOI: http://dx.doi.org/10.1136/jamia.2009.001024

[7] M. Ong, F. Magrabi,E. Coiera. Automated categorization of clinical incident reports using statistical text classification. Qual Saf Health Care 19 (2010) 1-7.
[8] M. Ong, F.Magrabi, E.Coiera. Automated identification of extreme-risk events in clinical incident reports. J Am Med Inform Assoc. 19 (2012) 110-8
[9] WEKA http://www.cs.waikato.ac.nz/ml/weka/downloading.html
[10] M. A. Hall. Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand. 1998.