MATEMATIKA, 2013, Volume 29, Number 1, 109–118

Comparison of the Naive Bayes Classifier and Instance Based Learner in Classifying Upper Gastrointestinal Bleeding

[1]**Nazziwa Aisha,** [2]**Mohd Bakri Adam and** [3]**Shamarina Shohaimi**

[1,2]Department of Mathematics, Faculty of Science
Universiti Putra Malaysia

[3]Department of Biology, Faculty of Science
Universiti Putra Malaysia

e-mail: [1]aishanazziwa@yahoo.ca.

Abstract Upper gastrointestinal bleeding is a medical emergence that results in high medical costs and death. Management of this disease requires ascertaining the cause of bleeding. The cause of bleeding is classified into esophageal and gastric causes. Based on health survey data, this study compares the accuracy of the naive Bayes classifier and an instance based learner in the classification of the cause of bleeding. The two classifiers are learned and trained on data collected from patients admitted for upper gastrointestinal bleeding. The naive Bayes classifier achieves a classification accuracy of 71% accuracy compared to 68% of the instance based learner.

Keywords Naive Bayes classifier, Gastrointestinal bleeding, IBk classifier, Peptic Ulcer disease.

2010 Mathematics Subject Classification Primary 62C10 secondary 62P10

1 Introduction

Non Variceal upper gastrointestinal bleeding (UGIB) is a common and challenging emergency for general physicians. It normally necessitates admission to hospital for urgent diagnosis and management. UGIB has many causes. These include esophageal causes like esophageal varices, esophagitis and esophageal cancer and gastric causes like gastric or peptic ulcer, gastric cancer and gastritis. The most common cause of UGIB is peptic ulcer disease (PUD) accounting for 20% of cases [1]. PUD is predominant in eldery with 68% over the age of 60 years and 27% over 80 years. The risk of bleeding is increased by some medications, like nonsteroidal anti-inflammatory drugs (NSAIDs). The highest risk is among people who require long-term use of very high-dose NSAIDs, especially patients with rheumatoid arthritis [2].

A meta analysis reported history of melena, hematemesis, ratio of Blood Urea Nitrogen/creatinine ratio (BUN) greater than 30 as predictive factors for a bleed coming from upper gastrointestinal source [3]. Other studies, however did not find BUN significant [4]. Hematemesis and melena i.e vomiting of red blood is the commonest presenting sign in UGIB [5]. In the absence of hematemesis, UGIB is likely if two or more of these factors are present: BUN ≥ 30, black stool and age is less than 50 years. A nasogastric aspirate is used to determine the source of bleeding if the factors are not present. When the aspirate is positive, a UGIB is greater than 50%, but not high enough to be certain and if it is negative then the source is likely to be lower. In this paper, we classify the the cause of UGIB into two; gastric or not gastric based on whether UGIB is caused by gastric or not. We develop a naive Bayes classifier (NBC) and an instance based classifier (IBk) based on the risk factors and presentation of the disease. The two classifiers are then used to determine

the cause of UGIB in the patients. Our objective is to determine which of the two classifiers is better in determining the cause of UGIB. The better classifier is one which achieves a higher clasification accuracy. The rest of the paper is organized as follows. The NBC and IBk classifier are explained in Section 2. The data set used and the methods of treating missing values in it are explained in Section 3. The methodology, results and conclusion are in Section 4, 5, 6 respectively.

2 Classifiers

Classification is the task of identifying to which of a group of categories a new observation belongs to, on the basis of a training set of data which has instances (or observations) with known categories. These observations have a set of properties that may be categorical or continuous. Some algorithms only work with discrete variables and they require that continuous variables be discretised. Classifiers can be established in three ways.

(i) Model the probability of class memberships given input data. e.g perceptron with the cross-entropy cost.

(ii) Make a probabilistic model of data within each class e.g. naive Bayes, model based classifiers.

(iii) model a classification rule directly e.g decision trees.

(i) and (iii) are discriminative classifiers, (ii) is a generative classifier and (i) and (ii) are probabilistic classifiers. Given that $C = C_1, \ldots, C_L, X = (X_1, \ldots, X_n)$, the discriminative probabilistic model is the model $P(C|X) = P(c_1|x)P(c_2|x), \ldots, P(c_L|x)$. The generative probabilistic model is the model $P(X|C)$. The generative probabilistic model for class 1, 2 and L is $P(x|c_1), P(x|c_2)$ and $P(x|c_L)$ where $x = (x_1, x_2, \ldots, x_n)$.

2.1 Naive Bayes Classifier (NBC)

The NBC discussed by Jensen [6] is a probabilistic classifier based on applying the Bayes theorem with strong independence assumptions. It assumes that the presence of a particular variable of a class is not related to the presence or absence of any other variable given the class variable. For example a patient may be diagnosed to have UGIB if he is over 60 years, has hematemesis, and BUN is greater than 30 mmol/L. Even if these variables depend on each other, a NBC considers all these variables to independently contribute to the probability that the patient has upper gastrointestinal bleeding.

2.1.1 The Probabilistic Model

A probability model for a classifier is a conditional model $P(C|X_1, \ldots, X_n)$ over a dependent class variable C with a small number of classes or outcomes conditional on a number of variables $X_1, \ldots, X_n$. For an instance to be classified, the NBC uses the Bayesian formula to calculate the probability of each class C given the values X_i of all the attributes. Assuming conditional independence of the attributes: X_i is conditionally independent of every other

attribute X_j for $j \neq i$. The joint model can be expressed as:

$$P(C|X_1, \ldots, X_n) = P(C) \prod_i^n P(X_i|C).$$

The NBC combines this model and the decision rule. The most common rule is the hypothesis that is most probable known as maximum a posteriori or MAP decision rule. The corresponding classifier is the function classify defined as

$$\text{classify}(f_1, \ldots, f_n) = \arg\max_c P(X = x) \prod_{i=1}^n P(X_i = x_i|C = C).$$

The NBC performs well, even when independence assumptions are clearly violated [7]. Since all attributes are used in determining those supporting a diagnosis and which ones are not, physicians find it easy to understand the way the NBC solves its tasks. The advantage of a NBC in modelling clinical data sets include:

(i) Its graphical nature makes it easy for clinicians to understand it as direct relationships between disease and its causes are represented.

(ii) They can be used to predict a target variable even with lots of uncertainty.

(iii) Given any subset of features observed, one can predict an output.

The naive Bayes has shown remarkable ability to outperform most advanced and sophisticated algorithms in many medical and also non-medical diagnostic problems.

2.1.2 Naive Bayes Algorithm(for Discrete Input Variables)

Learning phase: Given a training set S, for each target value of $c_i (c_i = c_1, \ldots, c_L)$,

$$\hat{P}(C = c_i) \longleftarrow \text{estimate} P(C = c_i) \text{ with examples in S.}$$

For every attribute value x_{jk} of each attribute $X_j (j = 1, \ldots, n; k = 1, \ldots, N_j)$

$$\hat{P}(X_j = x_{jk}|C = c_i) \longleftarrow \text{estimate} P(X_j = x_{jk}|C = c_i) \text{with examples from S.}$$

Output conditional probability tables; for $X_j, N_j * L$ elements.
Testing phase: Given an unknown instance $X = (a_1, \ldots, a_n)$ assign the lable c^* to X if

$$[\hat{P}(a_1|c^*), \ldots, \hat{P}(a_n|c^*)]\hat{P}(c^*) > \hat{P}(a_1|c), \ldots, \hat{P}(a_n|c)]\hat{P}(c), c \neq c^*$$

2.2 Instance Based Classifier (IBk)

The task of classification is carried out in two steps. The first is one of learning/training and the second is prediction/classification. During learning, a set of labeled data, called the training set, is used to learn the function which maps observations to classes. In the prediction step, the classifying function, learned during the training phase, is used to predict the classes of new data for which the appropriate class is unknown.

For the IBk classifiers, the algorithm stores the feature vectors and class labels during training. In the classification phase, an unlabeled vector is classified by assigning the label which is most frequent among the k training samples nearest to that query point (majority voting). One disadvantage with majority voting is that the classes with the most frequent examples will dominate the prediction of the new vector as they tend to come up in the k nearest neighbours when the neighbours are computed due to their large number [8]. This problem is overcome by weighing the classification while taking into account the distance from the test point to each of its k nearest neighbor.

2.2.1 Parameter Estimation

The best choice of k depends on the data. The accuracy of the k-NN algorithm is severely affected by presence of noisy and irrelevant features. Large values of k reduce the effect of noise on the classification, but make the boundaries of the class less distinct [9]. A good k can be selected by different heuristics techniques like cross validation. When $k = 1$, the class is predicted to be the class of the nearest training example and this is called the nearest neighbor algorithm.

2.2.2 The Basic Framework

The two phases i.e learning and prediction of an IBk can be formalized as follows; consider a set F of possible observations and a set C of possible assignment classes. $C = (c_1, c_2, \ldots, c_k) \subset N$. A training set consists of pairs (observations, class) belonging to the cartesian product $F \times C$. If we consider a set F, composed of feature vectors of cardinality L then a training set of cardinality n is of the type

$$FS \equiv \{(x_1^{(1)}, \ldots, x_1^{(L)}; c_1), \ldots, (x_n^{(1)}, \ldots, x_n^{(L)}; c_n)\} \in F \times C$$

with $k \leq n$. In the first phase, the generic learning algorithm (LA) defines a function of the type: $FS \in F \times C \rightarrow RI \in F \times C$ where RI is a possible representation of a learnt set. In the second phase, RI is used to classify/ predict classes of new observations. Observations are assigned classes with the function: $X \in F \rightarrow C_k \in C \subset N$. For a more detailed description of the IBk classifiers and its usage in medical studies as second-opinion diagnostic tools, see [10].

3 Data

The data consists of 480 cases of UGIB. Among these we have 120 patients with bleeding peptic ulcer, the commonest cause of gastrointestinal bleeding [11]. This data was obtained from one hospital in Malaysia. Twenty four variables consisting of risk factors and symptoms, were recorded for each case which include, age, Blatchford score total, blood urea nitrogen score, gastric malignancy, gastric polyps, gender, haemoglobin score for male and female, NSAID usage, pulse rate, systolic blood pressure, hematemesis and H2 liver failure or chronic liver. A quick cluster analysis enabled us eliminate some of the variables from further analysis which included Mallory-weiss tear, mucosal erosive disease, patient on anticoagulant, platelet or bleeding disorders, esophagitis, portal hypertension gastropathy, presentation type, Proton Pump Inhibitors (PPI). Age was grouped into two categories,

above and below 60. The reason for this is that in the literature it was found that people above 60 years of age are more likely to get peptic ulcer disease than the ones younger [12,13]. The characteristics of the data are shown in Table 1.

Table 1: Characteristics of the Data in the Population (n=480)

Variables	Categories	%
Gender	Male	81
	Female	19
Age	60 and Below	45
	Above 60	55
Blood Urea Score Range	State 1: $(< 6.5) mmol/L$	6
	State 2: $(6.5- < 8) mmol/L$	28
	State 3: $(8- < 10) mmol/L$	19
	State 4: $(10- < 25) mmol/L1$	33
	Missing	14
Gastric Malignancy	Yes	2
	No	98
NSAIDs Usage	Yes	76
	No	24
Pulse Rate(X), beats/min	Above 100	35
	below or 100	65
Hematemesis	Yes	21
	No	79
Peptic Ulcer Disease	Yes	27
	No	73

3.1 Parameter Estimation

All the model parameters were approximated with relative frequencies from the training set. Class priors were calculated as estimates for the class probability from the training set i.e (prior for class=yes is (number of patients whose cause of UGIB is gastric)/ (total number of sample). The parameters for the feature distribution were estimated by generating non parametric models from the training set. All continuous variables were discretised using binning.

Since the NBC uses all attributes available in its computations, it is very common that the data set will have incomplete cases. Like many data mining techniques, the NB are developed with the assumption that the data is complete and unfortunately this is not always the case. Typical data sets contain missing values either in the dependent or predictor variables.

3.2 Missing Data Treatment in This Paper

We included missing data treatments to investigate whether these missing data methods could provide information that would be important to improve the classification accuracy of the model. The variables that had missing values were haemoglobin score for male and female. We treat the missing values in five different ways. These ways are illustrated in Figure 1. In the figure A, B, C, D and E are the missing value treatments. Missing data after discretisation is shown in column 3.

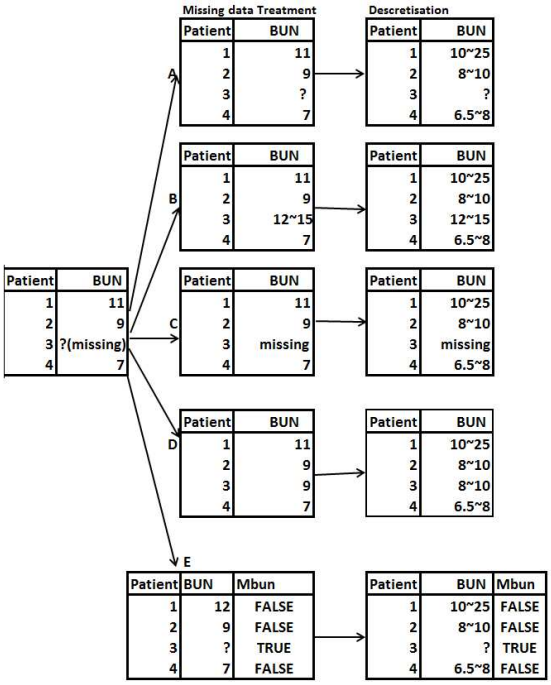

Figure 1: The Missing Data Treatments of the Variable Blood Urea Nitrogen (BUN) and its Discretisation

(A) Allow missing values i.e. no preprocessing the data, using Weka software [14]. Patient 3 has no results for Blood Urea Nitrogen (BUN). We leave the data as it is i.e no treatment of missing values.

(B) Substitute with normal values. In this method, a missing value is replaced with the known normal value of that variable, e.g, the normal haemoglobin score for female was 12 to 16 gm/dl(gram per deciliter), and for male was, 14 to 18 gm/dL [15]. These are the normal values that we substituted for the missing values according to this treatment. According to Lin and Haug [16], there are two reasons for missing clinical data.

   (a) A preference for reporting present symptoms over absent symptoms.

(b) Reporting more severe symptoms before those less severe.

A missing value suggests that a symptom was absent. In learning parameters missing values are stated as absent for discrete and for continuous variables, a missing value is assigned a typical value of a normal patient elicited from the expert. e.g if patient results for temperature were missing, we replace it with normal body temperature of 37 degrees Celsius.

(C) Give the missing values as a different level (missingness stratification approach). Here we assume, that "missing" is a state in itself.

(D) Impute the missing value with the overall mean or mode of the available values in the data set. The missing value is replaced with the mean which is calculated from

$$\frac{11+9+7}{3} = 9.$$

(E) Create an additional variable to represent missingness for each existing variable that was found to be absent in one or more patients (missingness indicator approach). Here we have created a variable mBUN which is missing BUN results. All instances with BUN present are indicated as false for the variable mBUN while all patients without BUN results are indicated as true for the variable mBUN.

4 Methodology

For each of the different missing data treatments, we learn a NBC and IBk classifier for our data. We use 10-fold cross validation to determine the classification accuracy of the classifiers [17]. In the 10-fold cross-validation procedure, the original data is randomly partitioned into 10 subsamples. Of the 10 subsamples, 9 subsamples are used as the training data set to create a model and the remaining single data is used for testing the model. The cross-validation process is then repeated 10 times, with each of the 10 subsamples used only once as the validation data. The results from the 10 folds are then averaged to produce a single estimation. All experiments were carried out using WEKA environment [18].

5 Result

Figure 2 shows the naive Bayes classifier developed to predict the cause UGIB with its initial prior probabilities. The class variable is gastric. It has two states: (yes, no) depending on whether upper gastrointestinal bleeding is caused by gastric or not gastric. The class variable is connected to all other variables and there are no other connections between the variables.

The probabilities are shown in percentage. The resulting percentage of correct classification are shown in Table 2. The NBC achieves classification accuracy of 71.04% compared to IBk with 68.5% when there is no treatment for missing data i.e (Treatment A). Treatment B and C improved accuracy of the NBC slightly while D and E decreased the accuracy. The decrease in accuracy when treatment D (Substitution with the mean) is used may be due to inserted bias by the mean. Mean imputation has also been found to bias results in other

studies [19]. For treatment E (Creating an additional variable), the decrease in accuracy could be due to the fact that more parameters have to be estimated due to an increase in the number variables. The IBk was not improved at all with the treatments of missing data. This raises a question as to whether these classifiers are better off with missing values than imputing. Studies have shown that the naive bayes classifier is less sensitive to missing data and can successfully be learned on data with missing values of up to 90% [20].

The classifiers developed can be used to find the probability given some observations about the patient. Table 3 shows the probability of the bleeding being caused by gastric when some features have been observed e.g, If a patient is above 60, and his BUN greater than 25mmol/L then the probability is 0.339 that UGIB in the patient is caused by a bleeding peptic ulcer. From Figure 2, the initial probability of having gastric was 0.265. When some observations about the patient are observed, the probability of gastric rose to 0.339.

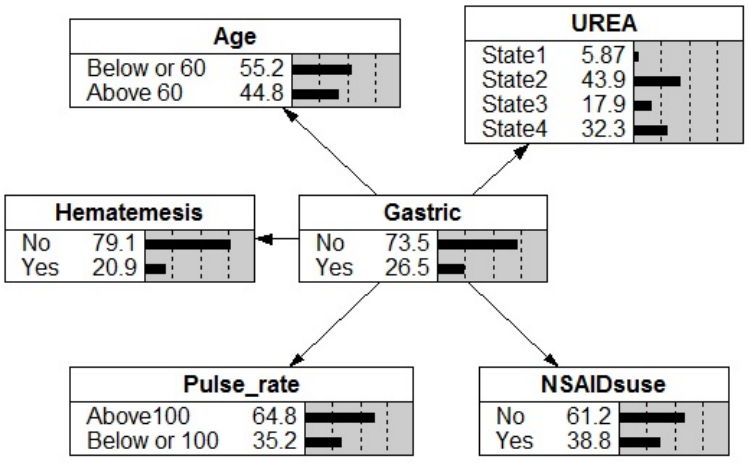

Figure 2: Naive Bayes Classifier for Upper Gastrointestinal Bleeding with Initial Probabilities

Table 2: Percentage of Correct Classification

Treatments	NBC	IBk
A	71.04	68.54
B	71.25	67.91
C	71.45	67.91
D	69.79	66.87
E	70.83	68.54

Table 3: Posterior Probability with Some Observed Features

Observed features		Probability
Age > 60	BUN $> 25 mmol/L$	No = 0.661 Yes = 0.339
Age ≤ 60	BUN $> 25 mmol/L$	No = 0.726 Yes = 0.274
Pulse Rate ≤ 100	BUN $(8- < 10\, mmol/L)$	No = 0.613 Yes = 0.387
Pulse Rate ≤ 100	BUN $(< 6.5\, mmol/L)$	No = 0.994 Yes = 0.006

6 Conclusion

We compared the classification accuracy of the naive Bayes classifier and the instance based learner on classifying the cause of upper gastrointestinal bleeding. The NBC perfomed better than the IBk classifier. Missing data in this paper was treated using five different methods. For all missing data treatments, the naive bayes classifier perfomed better than the IBk classifier. The IBk classifier was not improved by any of the missing data treatments. There are many methods of handling missing data that have not been explored in this paper like method of expectation maximisation. Further studies can explore these missing data methods to determine which of them will improve the classification accuracy of the NBC and the IBk classifier and to find out which of the two classifiers is better.

6.1 Acknowledgments

The author would like to thank the Ministry of Higher Education Malaysia for the scholarship.

References

[1] Boonpongmanee, S., Fleischer, D. E., Pezzullo, J. C., Collier, K., Mayoral, W., Al-Kawas, F., Chutkan, R., Lewis, J. H., Tio, T. L., and Benjamin, S. B. The frequency of peptic ulcer as a cause of upper-GI bleeding is exaggerated. *Gastrointestinal endoscopy.* 2004. 59(7):788–794.

[2] Huang, J. Q., Sridhar, S. and Hunt, R. H. Role of Helicobacter pylori infection and non-steroidal anti- inflammatory drugs in peptic-ulcer disease : a meta-analysis. *lancet.* 2002. 359: 14–22.

[3] Srygley, F. D., Gerardo, C. J., Tran, T. and Fisher, D. A. Does This Patient Have a Severe Upper Gastrointestinal Bleed? *JAMA: The Journal of the American Medical Association.* 2012. 307(10): 1072–1079.

[4] Al-Naamani, K., Alzadjali, N., Barkun, A. N. and Fallone, C. Does blood urea nitrogen level predict severity and high-risk endoscopic lesions in patients with nonvariceal upper

gastrointestinal bleeding? *Canadian journal of gastroenterology = Journal canadien de gastroenterologie*. 2008. 22(4): 399–403. ISSN 0835-7900.

[5] Zaragoza, A. M., Ten, J. M. and Llorente, M. J. Prognostic factors in gastrointestinal bleeding due to peptic ulcer construction of a predictive model. *J. of Clin Gastroenterology*. 2008. 42(7): 786–790.

[6] Jensen F. V. *An introduction to Bayesian Networks*. 1996.

[7] Lewis, D. D. Naive Bayes at Forty: The Independence Assumption in Inform ation Retrieval. *Machine Learning ECML98*. 1998. (x): 4–15.

[8] Coomans, D. and Massart, D. L. Alternative k-nearest neighbour rules in supervised pattern recognition: part 1. k-nearest neighbour classification by using alternative voting rules. *Analytica Chimica Acta*. 1982. 136:15–27.

[9] Everitt, B, Landau, S., Leese, S. and Stahl, D. *Miscellaneous Clustering Methods, in Cluster Analysis. 5th Edition.* Chichester, UK: John Wiley & Sons, Ltd. 2011.

[10] Gagliardi, F. Instance-based classifiers applied to medical databases: diagnosis and knowledge extraction. *Artificial intelligence in medicine*. 2011. 52(3):123–39.

[11] Palmer, K. Haematemesis and melena. *Medicine*. 2009. 37(1): 35–41. ISSN 13573039.

[12] Tanner, N. C. Surgery of peptic ulceration and its complications. *Postgraduate medical journal*. 1954. 30(349):577–592.

[13] Vonbach, P, Reich, R., Möll, F., Krähenbühl, S., Ballmer, P. E. and Meier, C. R. Risk factors for gastrointestinal bleeding: a hospital-based case-control study, *Drug-Drug Interactions in the Hospital*. 2007. 103.

[14] Bouckaert, R. R. *Bayesian Network Classifiers in Weka for Version 3-5-6*. The University of Waikato. 2007.

[15] Alvarez, G., Hébert, P. C. and Szick, S. Debate: transfusing to normal haemoglobin levels will not improve outcome, *Critical Care-London-*. 2001.5(2):56–63.

[16] Lin, J. H. and Haug, P. J. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *Journal of biomedical informatics*. 2008. 41(1): 1–14. ISSN 1532-0480.

[17] Kohavi, R. (1995). A study ofcross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, 14: 1137–1145.

[18] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

[19] Booth, D. E. Analysis of incomplete multivariate data. *Technometrics*. 2000. 42(2): 213–214.

[20] Liu, P., Lei, L., and Wu, N. A quantitative study of the effect of missing data in classifiers. *Computer and Information Technology, 2005. CIT 2005. The Fifth International Conference on*. IEEE. 2005. 28–33.