Basic Research Journal of Medicine and Clinical Sciences ISSN 2315-6864 Vol. 2(7) pp. 75-79 August 2013
Available online http//www.basicresearchjournals.org

Full Length Research Paper

Classification models for predicting the source of gastrointestinal bleeding in the absence of hematemesis

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

[1,2]Department of Mathematics, Islamic University in Uganda.
[3]Department of Biology, Faculty of Science, Universiti Putra Malaysia 43400 Serdang, Malaysia.

*Corresponding author email: aishanazziwa@yahoo.ca

Accepted 19 August, 2013

Management of acute gastrointestinal bleeding necessitates the identification of the source of bleed. The source of bleeding which is clear in patients presenting with hematemesis, is unclear in the absence of it. Logistic regression, decision tree, naïve Bayes, LogitBoost and KNN models were constructed from non endoscopic data of 325 patients admitted via the emergence department (ED) for GIB without hematemesis. The performance of the models in predicting the source of bleeding into upper gastrointestinal bleeding or lower gastrointestinal bleeding was compared. Overall the models demonstrate good performance with regards to sensitivity specificity, PPV, NPV and classification accuracy on the simulated data. On the GIB data, the naive Bayes model performed best with a prediction accuracy and sensitivity of 86%, specificity of 85% and area under curve of 92%. Classification models can help to predict the source of gastrointestinal bleeding for patients presenting without hematemesis and may generally be useful in decision support in the ED. The models should be explored further for clinical relevance in other settings.

Keywords: Upper Gastrointestinal bleeding, Emergency department, Hematemesis, Classification models, Naïve Bayes, Weka, Naive Bayes classifier, Random Forest, Decision tree, Logit boost, k nearest neighbour, Logistic regression.

INTRODUCTION

Acute upper gastrointestinal bleeding (UGIB) is a common medical emergence with 50-150 per 100,000 people admitted per year (Blower et al., 1997). It is predominant in elderly patients and is significantly higher in patients who are already admitted in hospital for co-morbidity or those with failure of endoscopic intention to treatment (Sostres and Lanas, 2011). According to Van Leerdam (2008), mortality in UGIB patients ranges between 3 and 14% and did not change in the past 10 years. In contrast to UGIB, mortality for Lower gastrointestinal bleeding (LGIB) during hospitalisation is very low (Hreinsson et al., 2013) and explains around 20% of all the cases of acute gastrointestinal bleeding (Zuccaro, 1998) with an incidence of around 20 per 100,000 populations per year in Westernized population (Longstreth, 1997).

In the emergency department, physician's make several decisions based on their prediction of the source of bleed (UGIB OR LGIB). The source of bleed will determine the consultant to be assigned and the timing for consultation. UGIB is diagnosed by an esophagoduodenoscopy while diagnosis of LGIB may also necessitate consultation with a general surgeon or nuclear medical specialist. The source of bleed may be determined by its manifestation. Hematemesis (vomiting of red blood), clearly shows that the bleeding is from the upper gastrointestinal tract (Gado et al., 2012) usually from an arterial source of varix. Earlier studies centred on patients with and without hematemesis have shown that the presence of hemodynamic instability, blood urea nitrogen to creatinine ratio, colour of stool, age, sex, may assist in the prediction of the source of GIB (Srygley et al., 2012), (Barkun et al., 2003). These attributes alone, in contrast with hematemesis are not diagnostic of upper or

lower GIB. E.g. melena which is black tarry stool is characteristic of UGIB but it may also indicate bleeding from the small bowel or right colon (Henneman, 2009). Other bleeding like chronic occult bleeding is detectable by chemical testing of a stool specimen and intra operative enteroscopy and yet the bleeding can occur anywhere in the gastrointestinal tract (Rockey, 1999). Currently, the nasogastric aspiration and lavage is a common procedure done in all patients with suspected UGIB to localize bleeding but it has low sensitivity and poor negative likelihood ratio, which limits its utility in ruling out an upper GI source of bleeding in patients with melena or hematochezia without hematemesis (Palamidessi et al., 2010).

A reliable classification model would be much helpful in identifying the source of GIB in patients without hematemesis. Such models have been constructed to predict acute upper gastrointestinal haemorrhage in patients with hematemesis (Chu et al., 2008), but have not been explored in the identification of the source of bleeding in patients presenting without hematemesis. In this study, we compare several classification models in the prediction of the source of GIB in a group of patients presenting without hematemesis. We use non endoscopic information that is available to emergence department (ED) physicians at the time of triage.

METHODS

Participants and study design

The study was a retrospective cohort study of GIB patients who were admitted through the ED. These patients were followed until they were discharge from the hospital. The study involved patients who visited two hospitals between January 1997 and December 2002: an academic tertiary care centre and a university affiliated community hospital were involved in the study. From the hospitals medical records department, patients, whose admitting diagnosis was related with the International Classification of Diseases, Ninth Revision codes related to GIB were obtained. The patient selected were those with the following: (1) were admitted via the ED for a principal diagnosis of GI tract bleeding, (2) heavy bleeding, as indicated by bloody or hemoccult positive black stools, or hemoccult positive dark stools if NGA was performed in the ED, (3) underwent confirmatory diagnostic testing within 3 days after admission, and (4) age 17 years or older. Exclusion criteria were: (1) hematemesis, (2) ostomy, (3) an obvious anorectal source, such as haemorrhoids and (4) admission for GI tract bleeding within the previous month. The protocol was approved by the institutional review board at each participating hospital. This study has been described in detail previously in (Witting et al., 2006).

Models and statistical analysis

Six models were trained to predict source of bleeding in patients presenting without hematemesis and their performance compared. These models are discussed briefly here.

- Naive Bayes (NB): Is a classifier based on Thomas Bayes theorem discovered in the 18th century (Jensen, 1996). The NB classifier assumes a conditional independence among the variables given the class variable. It first partitions the dataset into several sub datasets by the class label. Then in each sub dataset the maximum likelihood estimator, is obtained. When used for classification, NB predicts a new data point as the class with the highest posterior probability. This model has previously been used in medical studies (Medhekar et al., 2013).
- Logistic regression: This is a regression model that fits the log odds of the dependent to a linear combination of the independent variables. It is used mainly for binary responses, but there are extensions for multiple responses called multinomial logistic regression. The likelihood function can be maximized using numerical methods like Newton Raphson algorithm to obtain the coefficients. Logistic regression is widely used in medical studies (Chu et al., 2008) because it can be clearly and succinctly represented but it might not however, be able to produce complex models, leading to under fitting.
- Decision tree (J48): Is extensively described in (Quinlan, 1993). It is an algorithm that uses a set of examples which are already divided into classes. Each example consists of a set of attributes which can be symbolic or numeric. The algorithm chooses the attribute which divides the examples into their classes and partitions the data in a manner corresponding to the values of the attributes. This process is recursively used on each partitioned subset until all examples in the current subset have the same class. The results are represented as a tree with each node specifying an attribute. The classes are the terminal nodes of the tree and they correspond to sets of examples in which no more attributes are available. The decision tree has been used to predict antimicrobial activity of synthetic peptides (Lira et al., 2013).
- Random forest (RF): Is an ensemble based method developed by Breiman (2001) where a forest of classification trees is grown. Each individual classification tree outputs a class and the final output by the RF is the mode of the classes output by the individual trees. During training, a subset of the original data samples is randomly selected with replacement, to grow each tree. At each node on the tree, the best split for the node is determined from a random sample of all the variables. The number of variables chosen at the first node is the number of

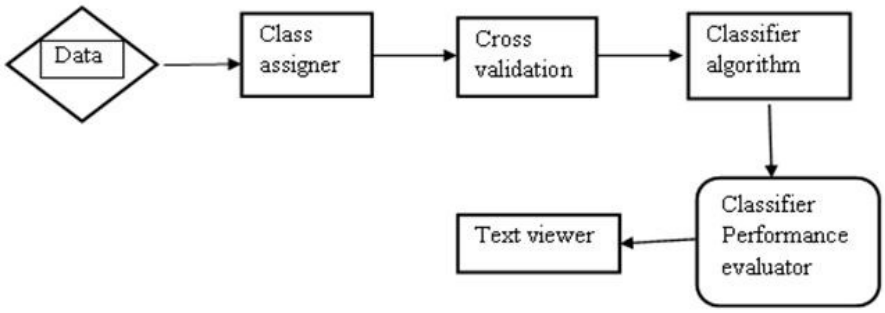

Figure1. Knowledge flow of the model building process.

variables selected for every node thereafter. All trees are grown to their fullest, with no pruning done. When the trees are grown, a test case is classified by majority voting among the trees. (Touw et al., 2013), gives a review of the use of the random forest in life science data.

- Logitboost: An ensemble voting method called boosting in which several weak classifiers are combined by weighted majority voting. When the weaker classifiers are combined, one more powerful and accurate classifier is produced. The Adaboost is one of the well known boosting algorithms. During the iterative process of the Adaboost, when a classifier is trained on a given iteration, and a wrong prediction is made for a particular case, this case is weighted more on the next iteration. At the end, we obtain a sequence of classifiers, with each new classifier learning from its mistakes. Finally a decision is made by majority voting among all classifiers. When the logistic regression is used as the cost function and the Adaboost as a generalized additive model the Logit Boost model is derived (Dettling, 2004). Previous studies have used the Logit boost in detecting coronary heart disease (Arsanjani et al., 2013).
- K-nearest neighbour (KNN). The KNN classifies a data point by considering the k closest neighbours to it. It is believed that the neighbours will be similar to each other. The Mahalanobis distance and Euclidean distance define the term "closest". While the Mahalanobis distance, considers the correlation of the data set and is scale invariant, the Euclidean distance. This model is simple to implement but faces a challenge in high dimensional data sets because neighbours' may not be nearby. Previous studies have applied the nearest neighbour concept for the prediction of protein secondary structure (Yang et al., 2013).

All the models described are available in the WEKA environment (Hall et al., 2009). Figure 1 shows the knowledge flow / model building process using WEKA software.

The data for use in the modelling procedure is entered into the software in .arff format or .csv. The variable \class that is to be predicted is assigned. In our study, the class variable is the source of bleeding i.e. UGIB or LGIB. The data is then split into ten parts using cross validation method. At the step of classifier algorithm, all the models to be investigated are included, i.e NB, RF, KNN, Logistic regression, J48 and logit boost. This enables evaluation and testing of a number of algorithms at the same time. The final output is shown at the text viewer.

Model Evaluation

We performed 10-fold cross validation for each iteration to obtain results with low mean square error (MSE) and bias. During cross validation, the data is split into 10 sub datasets. 90% of the data is used for training and the remaining 10% is for testing. This is repeated till every sub dataset becomes a testing set. Accuracy (sum of correct prediction divided by total predictions), specificity, sensitivity, negative predictive value NPV, positive predictive value (PPV) and ROC curves of the six predictive models were obtained for every 10-fold CV. The results of the 10 repetitions of 10-fold CV were then averaged and presented for the six models.

Simulation study

To compare the performance of the models further, we tested the models on simulated GIB data. The simulated data is of two types. One where the variables are independent of each other and the other is of correlated data. We took sample size of 500 patients in both cases and maintained the distributions of the real GIB data ie. 60.9% of the patients had LGIB and 30.9% had UGIB. This distribution is the same as the distribution in the real GIB data. Ten-fold cross validation was done using same procedure as described before. Accuracy, Roc curves, Specificity, Sensitivity, NPV and PPV values were averaged together. We maintained default parameters for all models.

Table 1. Model performance with GIB data.

Model	ACC	SN	SP	NPV	PPV	ROC
NB	0.86	0.86	0.85	0.81	0.89	0.92
J48	0.81	0.81	0.77	0.75	0.85	0.85
LogitBoost	0.85	0.85	0.82	0.81	0.88	0.93
Logistic	0.85	0.85	0.82	0.79	0.88	0.92
KNN	0.83	0.83	0.74	0.80	0.85	0.87
RF	0.84	0.84	0.82	0.78	0.88	0.88

ACC=Accuracy, SN=Sensitivity, SP=Specificity, PPV= Positive Predictive Values, NPV= Negative Predictive Value

Table 2. Simulated study results (Independent GIB data).

Model	ACC	SN	SP	NPV	PPV	ROC
NB	0.90	0.94	0.84	0.91	0.89	0.96
J48	0.88	0.93	0.81	0.89	0.87	0.90
LogitBoost	0.91	0.95	0.85	0.92	0.90	0.96
Logistic	0.90	0.93	0.85	0.89	0.90	0.96
KNN	0.89	0.95	0.81	0.92	0.88	0.94
RF	0.90	0.93	0.85	0.89	0.90	0.93

ACC=Accuracy, SN=Sensitivity, SP=Specificity, PPV= Positive Predictive Values, NPV= Negative Predictive Value

Table 3. Simulated study results (Correlated GIB data).

Classifier	ACC	SN	SP	NPV	PPV	ROC
NB	0.85	0.90	0.77	0.80	0.87	0.92
J48	0.85	0.89	0.81	0.79	0.89	0.87
LogitBoost	0.86	0.89	0.80	0.80	0.89	0.93
Logistic	0.85	0.87	0.83	0.77	0.90	0.93
KNN	0.86	0.90	0.78	0.81	0.88	0.93
RF	0.85	0.89	0.79	0.79	0.88	0.91

ACC=Accuracy, SN=Sensitivity, SP=Specificity, PPV= Positive Predictive Values, NPV= Negative Predictive Value

RESULTS

Table 1 depicts the performance of each model for the response variable in the real data set and Table 2 and 3 show the model performance on independent simulated data and correlated data respectively. Overall the models are found to demonstrate good overall performance with regards to sensitivity specificity, PPV, NPV and accuracy. For the real data, accuracy obtained by the naive Bayes model was superior to others with the model correctly predicting the source of bleeding 86% of the time. The areas under the ROC curves for this model were 92%. The source of GIB was correctly predicted with 85% using Logistic regression and LogitBoost regression and with ROC curves of 92% and 93% of the time. On average, all models performed better on the simulated data than on original dataset. The NB however, did not perform well on the simulated correlated data. A possible reason for this is its failure to discover relationships between variables. This is because the NB is built on the conditional independence assumption which states that variables are conditionally independent given the class variable. Other studies have also confirmed the poor performance of NB on such data (Domingos and Pazzani, 1996). Generally, the NB performs well even in cases of violations of this assumption.

DISCUSSION AND CONCLUSION

For predictive models to be of use in decision making in the emergence department, an important feature is that they must be able to use data that is quickly available to the clinician at the time of admission. The six models were built based on data obtained from physical examination, clinical history, and initial laboratory investigation. We emphasize that identification of the source of bleeding is best done by a gastroenterologist and that our models are not meant to replace an experienced clinician. Our models performed well with

accuracies exceeding 80%. This finding suggests that these models may be likely to identify the source of GIB in patients presenting without hematemesis. These patients could be easily assigned to specific physicians for diagnosis and further management. For our study the naive Bayes and LogitBoost performed well in agreement with previous studies.

Logistic regression is a widely accepted and used model in medical studies. In our study, it was competitive, with the NB model. The NB has an advantage over the logistic regression model, in a way that it is able to use more input variables than the logistic regression based models and their fore the interaction between many clinical variables can be investigated.

Our results also support the conclusion by Plant and Böhm (2010), that J48 achieves less accuracy in some medical datasets. In our study it had the lowest performance in all cases. Unlike the study by Chu et al. (2008) which found the RF having the highest accuracy, the RF was not the best model in this study. RF performs well on high dimensional datasets with large number of features compared with sample size. Our study had less number of features and a larger sample size compared to Chu et al. (2008). Although both studies are on prediction of source of GIB, our study is unique in a way that we dealt with patients who do not present with hematemesis which is the most important factor in the prediction of UGIB.

Although classification models have performed well in many medical studies, they have not been widely used and accepted by physicians. The barriers to the use of classification models by physicians need to be identified, so that they can be used in their overall decision making process.

REFERENCES

Arsanjani R, Vahistha V, Shalev A, Nakanishi R, Hayes S, Fish M, Berman D (2013). Improved accuracy of myocardial perfusion SPECT for detection of coronary artery disease by machine learning in a large population. J. Nuclear Cardiol. 1–10.

Barkun A, Bardou M, Marshall JK (2003). Consensus recommendations for managing patients with nonvariceal upper gastrointestinal bleeding. Annals of Internal Medicine, 139(10): 843–857.

Blower AL, Brooks A, Fenn GC, Hill A, Pearce MY, Morant S, Bardhan KD (1997). Emergency admissions for upper gastrointestinal disease and their relation to NSAID use. Alimentary pharmacology and therapeutics, 11(2): 283–91. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9146764

Breiman L (2001). Random forests. Machine learning, 45(1): 5–32.

Chu A, Ahn H, Halwan B, Kalmin B, Artifon ELA, Barkun A, Lagoudakis MG (2008). A decision support system to facilitate management of patients with acute gastrointestinal bleeding. Artificial Intelligence in Medicine, 42(3): 1–41.

Dettling M (2004). BagBoosting for tumor classification with gene expression data. Bioinformatics, 20(18): 3583–3593.

Domingos P, Pazzani M (1996). Beyond independence: Conditions for the optimality of the simple bayesian classi er. Proc. 13th Intl. Conf. Machine Learning (pp. 105–112).

Gado AS, Ebeid BA, Abdelmohsen AM, Axon AT (2012). Clinical outcome of acute upper gastrointestinal hemorrhage among patients admitted to a government hospital in Egypt. Saudi J. gastroenterol. official journal of the Saudi Gastroenterology Association, 18(1): 34.

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

Henneman PL (2009). Gastrointestinal bleeding. Rosen's Emergency Medicine: Concepts and Clinical Practice. 7th ed. Philadelphia, Pa: Mosby Elsevier.

Hreinsson JP, Gumundsson S, Kalaitzakis E, Björnsson ES (2013). Lower gastrointestinal bleeding: incidence, etiology, and outcomes in a population-based setting. Euro.J. Gastroenterol. and hepatol. 25(1): 37–43.

Jensen FV (1996). An introduction to Bayesian Networks. UCL press London, England.

Lira F, Perez PS, Baranauskas JA, Nozawa SR (2013). Prediction of antimicrobial activity of synthetic peptides by a decision tree model. Applied and Environmental Microbiol. 79(10): 3156–3159.

Longstreth GF (1997). Epidemiology and outcome of patients hospitalized with acute lower gastrointestinal hemorrhage: a population-based study. American J. Gastroenterol. 92(3): 419–424.

Medhekar DS, Bote MP, Deshmukh SD (2013). Heart Disease Prediction System using Naive Bayes. Heart Disease, 2(3).

Navot A, Shpigelman L, Tishby N, Vaadia E (2005). Nearest neighbor based feature selection for regression and its application to neural activity. Advances in Neural Information Processing Systems (NIPS), 19, 995–1002.

Palamidessi N, Sinert R, Falzon L, Zehtabchi S (2010). Nasogastric aspiration and lavage in emergency department patients with hematochezia or melena without hematemesis. Academic Emergency Medicine, 17(2): 126–132.

Plant C, Böhm C (2010). Database technology for life sciences and medicine (Vol. 6). World Scientific Publishing Company Incorporated.

Quinlan JR (1993). C4.5: Programs for Machine Learning. Machine Learning (Vol. 240). Morgan Kaufmann. Retrieved from http://portal.acm.org/citation.cfm?id=152181

Rockey DC (1999). Occult gastrointestinal bleeding. New England Journal of Medicine, 341(1): 38–46.

Sostres C, Lanas A (2011). Epidemiology and demographics of upper gastrointestinal bleeding: prevalence, incidence, and mortality. Gastrointestinal endoscopy clinics of North America, 21(4): 567.

Srygley FD, Gerardo CJ, Tran T, Fisher DA (2012). Does this patient have a severe upper gastrointestinal bleed? JAMA: The J. the American Medical Association, 307(10): 1072–1079.

Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, Van Hijum SAFT (2013). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Briefings in bioinformatics, 14(3): 315–326.

Van Leerdam ME (2008). Epidemiology of acute upper gastrointestinal bleeding. Best Practice & Research Clinical Gastroenterology, 22(2); 209–224.

Witting, M. D., Magder, L., Heins, A. E., Mattu, A., Granja, C. A., & Baumgarten, M. (2006). ED predictors of upper gastrointestinal tract bleeding in patients without hematemesis. The American journal of emergency medicine, 24(3), 280–285.

Yang W, Wang K, Zuo W (2013). Prediction of protein secondary structure using large margin nearest neighbour classification. Intern. J. Bioinformatics Res. appl. 9(2): 207–219.

Zuccaro G (1998). Management of the adult patient with acute lower gastrointestinal bleeding. The Amer. J. Gastroenterol.93(8): 1202–1208.