On the Use of Bayesian Network Classifiers to Classify Patients with Peptic Ulcer Among Upper Gastrointestinal Bleeding Patients.

Nazziwa Aisha[1] Mohd Bakri Adam[2]

[1,2]Department Mathematics.
Universiti Putra Malaysia
Serdang, Malaysia
e-mail aishanazziwa@yahoo.ca, bakri@science.upm.edu.my

Abstract— **A Bayesian network classifier is one type of graphical probabilistic models that is capable of representing relationship between variables in a given domain under study. We consider the naïve Bayes, tree augmented naïve Bayes (TAN) and boosted augmented naïve Bayes (BAN) to classify patients with peptic ulcer disease among upper gastro intestinal bleeding patients. We compare their performance with IBk and C4.5. To identify relevant variables for peptic ulcer disease, we use some methodologies for attributes subset selection. Results show that, blood urea nitrogen, hemoglobin and gastric malignancy are important for classification. BAN achieves the best accuracy of 77.3 and AUC of (0.81) followed by TAN with 72.4 and 0.76 respectively among Bayesian classifiers. While the accuracy of the TAN is improved with attribute selection, the BAN and IBK are better off without attribute selection.**

Keywords- Feature selection, Peptic ulcer disease, Bayesian network classifiers, Gastro intestinal bleeding, Classification

I. INTRODUCTION

Probabilistic graphical models, such as Bayesian networks and influence diagrams, are based on sound foundations of probability theory. They combine available statistics with expert judgment to represent domain knowledge under study. Bayesian networks have been successfully applied in medicine, for medical diagnosis [1][2]. A number of factors affect the success of the probabilistic graphical models. Among these is the presence of noisy, redundant and unreliable information. This information may lead to over fitting of the data and an increase in structural complexity. To generate networks that are simple to evaluate, reduce the data size and complexity and increase the overall accuracy of these models, we need to select a subset of attributes that are relevant for classification tasks [3]. We describe our work on a probabilistic model for classifying bleeding peptic ulcer. We address the problem of identifying a small subset of attributes for use during the classification. Our aim is to select a subset of variables while using some of the available feature selection techniques. We determine whether Bayesian Network classifier (BNC) performance is enhanced with attribute selection.

II. BAYESIAN NETWORK CLASSIFIER

A bayesian network is a directed acyclic graph that represents a probability distribution over a set of values *U*. A belief Network for *U* is a pair *B* = (G; B). The first component *G*, is a directed acyclic graph whose vertices correspond to the random variables $X_1,\ldots,X_n$ that encodes the following set of conditional independencies assumptions. Each variable X_i is independent of its non descendants given its parents in *G*. The second component of the pair B, represents the set of parameters, that quantifies the network. It contains a parameter $\theta x_i \,|(Pa_i) = P(x_i|(Pa_i))$ for each possible value of x_i of X_i and pa_i of Pa_i. Together with the graph structure, they are sufficient to represent the joint probability distribution of the domain [1], given *by* $P(X_{1,}...X_n) = \prod_{i=1}^{n} P(X_i \,|\, (Pa_i))$ where Pa_i is the set containing the parents of Xi in the Bayesian Network.

This network is easily adopted for classification tasks. Naïve Bayes (NB) classifier is a special case of Bayesian networks with the assumption that every attribute, is independent of the other attributes given the class variable. When this assumption is relaxed or removed we obtain the tree augmented naïve Bayes (TAN) and boosted augmented naïve Bayes (BAN) respectively [2].

II. DATA

Data on gastrointestinal bleeding was obtained from a big hospital in Sabah, Malaysia. Urea was discretised into 4 groups which include blood urea nitrogen score range (BUN*).* *(BUN) < 6.5millimoles per litre (mmo/L), 6.5mmol/L BUN < 8mmol/L, 8mmol/L ≤ BUN < 10mmol/L*, *10mmol/L ≤*

Nazziwa Aisha is sponsored by Ministry of Education Malaysia works with the Islamic University In Uganda. Mohd Bakri Adam is with Universiti Putra Malaysia

25mmol/L. Age and SBP were grouped into less than and above *60*, and *SBP < 110, SBP ≥ 110* respectively. Hemoglobin for female was grouped into two, below and above 10, and others while hemoglobin for male was grouped into less than *10gramms per deciliter (g/dL), 10g/dL to 12g/dL* and greater than *13g/dL*. All the other variables were binary having the values present or absent [3]. For further information about factors for acute upper gastro intestinal bleeding see [4]. All the work was done using Weka (Waikato environment for knowledge analysis). Weka is freely available from http://www.cs.waikato.ac.nz/hml.

III. ATTRIBUTE SELECTION TECHNIQUES

a) *Filter methods*

These methods are performed prior to learning, during preprocessing stage. They use general characteristics of the data to assess the relevance of the attributes removing low scoring attributes.

1) Information gain criteria *(IG)* .

Information gain is the additional information provided by the attribute that is shown by a reduction in entropy of the class [5]. Each attribute is given a score based on the Information gain between itself and the class. If A is an attribute and C is the class, the entropy before and after observing the class is given by:

$$H(C) = -\sum_{c \in C} p(c) \log_2 p(c),$$

and

$$H(C) = -\sum_{a \in A} P(a) \sum_{c \in C} p(c \mid a) \log_2 p(c \mid a)$$

2) *ReliefF*

This method randomly samples an instance from the data and then locates its nearest neighbor from the same and opposite class. The values of the attributes of the nearest neighbors are then compared to the sampled instance and used to update relevance scores for each attribute [6]. The number of instances that the process is repeated need to be specified.

3) *Consistence Based Subset Evaluation.*

These methods look for combinations of attributes whose values divide the data into subsets containing a strong single class majority. Usually the search is biased in favor of small feature subsets with high class consistence [7]. We use consistence metric where; Consistence (cs) is given by

$$cs = 1 - \frac{\sum_{i=0}^{j} |D_i| - |M_i|}{N}$$

Where s is an attribute subset, J is the number of distinct combinations of attribute values for s, Di is the number of occurrences of the i^th^ attribute value combination, Mi is the cardinality of the majority class for the i^th^ attribute value combination and N is the total number of instances in the data set.

4) *Correlation-Feature Selection.*

This is a subset evaluation heuristic. It evaluates subsets of attributes rather than individual attributes. Subsets of attributes that have low inter correlation with each other and are highly correlated with the class are given high scores. Low scoring subsets are removed [8].

$$\text{Merits} = \frac{k\overline{r_c f}}{\sqrt{k + k(k-1)\overline{rff}}}$$

where Merits is the heuristic merit of an attribute subset S containing k attributes, $\overline{r_c}$f the average attribute-class correlation, and rff the average feature-feature inter-correlation. The numerator indicates how predictive a group of features are and the denominator gives an indication of how much redundancy there is among them. This heuristic handles irrelevant features as they will be poor predictors of the class. In order to apply the above Equation we first compute the correlation (dependence) between attributes. CFS first discretizes numeric features and then uses symmetrical uncertainty to estimate the degree of association between discrete features (X and Y):

$$SU = 2[\frac{H(X) + H(Y) - H(X,Y)}{H(X) + H(Y)}]$$

When the correlation matrix is computed, CFS applies a heuristic search strategy to find a good subset of features according to the above equation.

b) *Wrapper methods*

These methods search through the space of all possible attribute subsets using estimated accuracy from a classification algorithm. The optimal subset of attributes is built into the classifier construction. Cross-validation is used to provide an estimate for the accuracy of a classifier on data when using only the attributes in a given subset [9].

c) *Embedded*

It's similar to the wrapper approach. In this set up, the search for an optimal subset of attributes is built into the classifier construction, and can be seen as a search in the combined space of attributes subsets and hypotheses. In this paper we focus on the C4.5 Algorithm [10].

IV. EXPERIMENTAL METHODOLOGY

The data set was divided into two thirds for training and one third for testing. We use the NB, TAN, the BAN, C4.5 and the IBK[11], algorithm, to compare the effectiveness of the attribute selection techniques. We obtained the percentage of correct classifications for each classifier, using all the 15 initial variables. The six attribute selection techniques are then applied to our data set. Using the selected attributes, we obtain the classification rate for each classifier.

For the C4.5 algorithm, we performed a fivefold cross validation and fixed to 10 the minimum number of instances per leaf, to provide an estimate of the accuracy of the classifier. Cross-validation is repeated as long as the standard deviation over the runs is greater than one percent of the mean accuracy or until five repetitions have been complete [12]. For all experiments we chose a threshold of 0.01

V. RESULTS AND CONCLUSION

Table 1 shows the number of attributes selected by each attribute section technique. The Accuracy, Specificity, Sensitivity, Negative Predictive Values (NPV) and Positive Predictive values (PPV) obtained by each classifier with the selected variables are shown in Tables 2A-2E. Since the wrapper selection methods are joined with the classifier, the number of attributes selected with each classifier is given and the classification rate is shown in Table 3.

IG selects the least number of attributes. This is also true in other studies [12]. Its selected attributes when used for classification produce poor ROC results. While consistence selects many attributes, it does not select gastric malignancy yet it is selected by the other three techniques. This pattern is also seen with the ReliefF which leaves out Fresh PR bleed. When all attributes are used in classification, IBK and C4.5 achieve the highest classification accuracy and area under curve (AUC) as shown in Table 2A.
Among the BNC, the BAN has the highest classification with correct accuracy of 77.3% and area under curve (AUC) of 0.81.

TABLE 1. ATTRIBUTES SELECTED

	Attribute selection techniques			
Attributes	IG	CFS	ReliefF	Consistence
Blood Urea Score	X	X	X	X
heamoglobin	X	X	X	X
Fresh PR Bleed	X	X		X
Gastric Malignancy	X	X	X	
NSAID's	X	X	X	
Liver failure		X		
Pulse Rate				X
Systolic Blood Pressure			X	X
Portal Hypertensive Gastropathy			X	X
Age			X	
Gender			X	
Hematemesis			X	
Gastric Polyps				

X indicates the attributes selected by the selection techniques

TABLE 2A. CLASSIFICATION WITHOUT ATTRIBUTE SELECTION

Algorithm	Ac%	Sp%	Se%	PPV%	NPV%	ROC
NB	67.8	88	40	19	88	0.71
TAN	72.4	74	56	24	93	0.76
BAN	77.3	78	71	39	93	0.81
C4.5	80.3	64	67	65	87	0.86
IBk	82.2	64	73	65	90	0.91

Ac= Accuracy, S= Specificity, Se= Sensitivity, PPV= Positive Predictive Values, NPV= Negative Predictive Value

TABLE 2B. CLASSIFICATION WITH IG SELECTED ATTRIBUTES

Algorithm	Ac %	Sp%	Se%	PPV%	NPV%	ROC
NB	71.5	71	71	3	100	0.72
TAN	66.6	75	42	36	79	0.67
BAN	70.6	71	0	0	100	0.50
C4.5	67.5	71	36	14	90	0.67
IBk	71.5	71	71	3	100	0.72

Ac= Accuracy, S= Specificity, Se= Sensitivity, PPV= Positive Predictive Values, NPV= Negative Predictive Value

TABLE 2C. CLASSIFICATION WITH RELIEFF SELECTED ATTRIBUTES

Algorithm	Ac%	Sp%	Se%	PPV%	NPV%	ROC
NB	*71.5*	*72*	*58*	*12*	*97*	*0.70*
TAN	*79.1*	*80*	*74*	*45*	*93*	*0.81*
BAN	*74.2*	*77*	*60*	*36*	*90*	*0.77*
C4.5	*80.0*	*84*	*69*	*59*	*87*	*0.86*
IBk	*71.5*	*72*	*58*	*12*	*97*	*0.70*

Ac= Accuracy, S= Specificity, Se= Sensitivity, PPV= Positive Predictive Values, NPV= Negative Predictive Value

TABLE 2D. CLASSIFICATION WITH CFS SELECTED ATTRIBUTES

Algorithm	Ac%	Sp%	Se%	PPV%	NPV%	ROC
NB	71.5	71	100	3	100	0.72
TAN	70.9	72	53	10	96	0.70
BAN	71.5	72	100	3	100	0.70
C4.5	70.6	71	0	0	100	0.50
IBk	69.6	75	48	33	83	0.74

Ac= Accuracy, S= Specificity, Se= Sensitivity, PPV= Positive Predictive Values, NPV= Negative Predictive Value

It is followed by the TAN with 72.4 and 0.76 respectively.The NBC is last with 67.8 and 0.71 respectively. While attribute selection improves the accuracy of the naïve Bayes classifier in all experiments, the accuracy of the BAN is decreased. The IBK is better off without attribute selection. This is because it is not improved by attribute selection. The C4.5 performed worse with less number of attributes selected.

TABLE 2E. CLASSIFICATION WITH CONSISTENCE SELECTED ATTRIBUTES

Algorithm	Ac%	Sp%	Se%	PPV%	NPV%	ROC
NB	68.4	70	31	6	94	0.66
TAN	70.2	71	46	7	96	0.68
BAN	72.4	72	58	2	93	0.72
C4.5	74.8	74	76	20	97	0.73
IBk	76.1	75	78	26	96	0.77

Ac= Accuracy, S= Specificity, Se= Sensitivity, PPV= Positive Predictive Values, NPV= Negative Predictive Value

TABLE 3. CLASSIFICATION RATE USING WRAPPER TECHNIQUE

	Algorithms				
	BAN	NB	TAN	C4.5	AODE
Number of Attributes	3	2	3	1	3
Accuracy %	77.91	73.54	73.54	73.54	77.91

Other studies, however found that it improves with attributeselection [13]. We noticed however, that the author uses genetic algorithm for attribute selection. We did not use this in our study. The wrapper improves performance on the BAN and AODE classifiers. The BNC performed better with less number of variables. The highly ranked attribute selected by all the techniques were gastric malignancy, Liver Failure, blood urea nitrogen and hemoglobin. These factors have been found significant in other studies [14].

We described our work on a probabilistic causal model for diagnosis of peptic ulcer disease. The model includes 15 attributes, such as important symptoms and signs. We found that, gastric malignancy, NSAID's usage, Fresh pr bleed, Liver failure, were important for classification of peptic ulcer disease. Blood Urea Score Range and hemoglobin score were also useful.

Given a patient's case, i.e. observation of values of any subset of the 14 attributes, the model computes the posterior probability distribution of having peptic ulcer disease. The class with the highest posterior probability indicates the class of a patient. There are many variables that are important in the prediction of peptic ulcer disease that were not available for the study.

REFERENCES

[1] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[2] Y. Jing, V. Pavlović, and J. M. Rehg, “Boosted Bayesian network classifiers,” *Machine Learning*, vol. 73, no. 2, pp. 155–184, 2008.

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

[4] G. F. Longstreth and others, “Epidemiology of hospitalization for acute upper gastrointestinal hemorrhage: a population-based study.,” *The American journal of gastroenterology*, vol. 90, no. 2, p. 206, 1995.

[5] H. Uguz, “A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm,” *Knowledge-Based Systems*, 2011.

[6] I. Kononenko, “Estimating attributes: Analysis and extensions of RELIEF,” *Machine Learning ECML94*, vol. 784, pp. 171–182, 1994.

[7] M. Dash and H. Liu, “Consistency-based search in feature selection,” *Artificial Intelligence*, vol. 151, no. 1–2, pp. 155–176, Dec. 2003.

[8] M. A. Hall and L. A. Smith, “Feature Subset Selection : A Correlation Based Filter Approach.” Springer, 1997.

[9] R. Kohavi and H. John, “Wrappers for feature subset selection,” *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273–324, 1997.

[10] J. R. Quinlan, “Induction of Decision Trees,” *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[11] D. W. Aha, D. Kibler, and M. K. Albert, “Instance-based learning algorithms,” *Machine learning*, vol. 6, no. 1, pp. 37–66, 1991.

[12] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” *international joint Conference on artificial intelligence*, vol. 14, pp. 1137–1145, 1995.

[13] M. Anbarasi, E. Anupriya, and S. N. IYENGAR, “Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm,” *International Journal of Engineering Science and Technology*, vol. 2, no. 10, pp. 5370–5376, 2010.

[14] C. Rollhauser, D. E. Fleischer, and others, “Nonvariceal upper gastrointestinal bleeding,” *Endoscopy*, vol. 34, no. 2, pp. 111–118, 2002.