Early Screening for Gastric Cancer Using Machine Learning Techniques

W.Z. Liu, A.P. White & M.T. Hallissey

Birmingham University, Edgbaston, P.O. Box 363, Birmingham B15 2TT,U.K. [1]

Abstract. The feasibility of using machine-learning techniques to screen dyspeptic patients for those at high risk of gastric cancer was demonstrated in this study. Data on 1401 dyspeptic patients over the age of 40, consisted of 85 epidemiological and clinical variables and a gold-standard diagnosis, made by upper gastrointestinal endoscopy. The diagnoses were grouped into two classes — those at high risk of having (or developing) gastric cancer and those at low risk. A machine-learning approach was used to generate a cross-validated sensitivity-specificity curve in order to assess the power of the discrimination between the two groups.

Acknowledgement. This work was partially supported by the Cancer Research Campaign under grant no. SP2229/0101.

1 Introduction

Gastric cancer is an extremely serious condition and those unfortunate enough to suffer from it stand little chance of survival unless a diagnosis is made and an operation performed at an early stage in the development of the disease. The problem is that, in the early stages of the condition, the patient suffers from a variety of dyspeptic symptoms which could easily indicate the presence of any of about a dozen different gastric complaints. The only way to be sure of the diagnosis is to make use of an endoscope to carry out an internal examination of the patient. Unfortunately, endoscopic examination of *all* dyspeptic patients is too expensive for the National Health Service to support. Therefore, the goal was to use machine learning techniques to identify a *subset* of dyspeptic patients at high risk of having gastric cancer, so that they could undergo endoscopic examination. The immediate objective was to generate a cross-validated sensitivity-specificity curve in order to assess the discriminative power of the technique.

2 Variable Classification Thresholds

Typically, machine learning software produces estimated posterior probabilities of class membership of new cases undergoing classification. These are obtained from

[1] W.Z. Liu is in the School of Mathematics and Statistics, where A.P. White is an Associate Member. M.T. Hallissey is in the Department of Surgery.

		Predicted Class	
		1	0
Actual Class	1	a (hit)	b (miss)
	0	c (FP)	d (CR)

Table 1: Frequency table for the classification outcomes in a two-class discrimination task. FP represents false positive and CR represents correct rejection.

the frequency counts of the cases in the training set found at the terminal nodes. In the field of medical diagnosis, it is often the case that a test for a particular disease does not operate with perfect accuracy. This leads to the possibility of two different types of error — false positive error in which the disease is predicted when it is actually absent and false negative error (or 'miss') where the test result is negative when the disease is present. It is clear that, in this application, the second type of error is more serious than the first. Consequently, the usual approach of classifying according to the largest posterior probability found at the terminal node is not satisfactory in such a situation. A more flexible approach to such situations is to adopt classification *threshold* probabilities. With the diagnosis of serious medical conditions, the classification threshold would typically be set at a lower value for predicting the presence of the disease, than for predicting its absence. Of course, the ability to vary the classification threshold means that, as the error rate for false negatives is decreased, the error rate for false positives increases. 'Hits' and correct rejections are similarly inversely related. Table 1 displays the four possible outcomes for a two-class discrimination task. From this table, some important quantities can be defined. In the terminology of medical diagnosis, *sensitivity* and *specificity* are defined as follows. Sensitivity is the conditional probability of correct classification, given that the disease is present and specificity is the conditional probability of correct classification, given that the disease is absent. Algebraically, from Table 1, sensitivity is given by $a/(a+b)$ and specificity is defined as $d/(c+d)$. As the classification threshold for the disease is reduced, sensitivity will *increase* and specificity will *decrease*, i.e. there is a trade-off between the two. If the classification threshold probability is varied from 0 to 1, values for sensitivity and specificity are generated as a series of number-pairs, which can be plotted on a graph. Another quantity of interest for two-class discrimination tasks is the *odds ratio.* From the foregoing table, this is defined as ad/bc. It is a measure of the magnitude of association in a 2×2 table. In the current context, it is useful because it can be used as a measure of discrimination power.

Sensitivity-specificity curves run from coordinates (0, 1) to (1, 0). A straight line corresponds to an odds ratio of one (i.e. absence of any effective discrimination). In general, a sensitivity-specificity curve will 'bulge' into the upper-right quadrant of the graph. The more extreme the curve, the larger the corresponding odds ratio (i.e. the better the discrimination). Sensitivity-specificity curves are useful because they show the relationship between these two quantities throughout the range of threshold values and, with the use of odds ratio contours, indicate the discrimination power at different threshold settings.

3 Method

Part of the database of dyspeptic patients described by Hallissey et al. (1990) was used for the study. Briefly, this consisted of records on 1401 patients over the age of 40 who were referred to dyspepsia clinics because of symptoms of dyspepsia.

The data on each patient comprised a gold-standard diagnosis (made by upper gastrointestinal endoscopy) and a total of 85 epidemiological and clinical variables. The diagnoses were grouped into two classes. Class 1 (the high risk group) consisted of those patients diagnosed as either having gastric cancer, or belonging to any of three other diagnostic categories regarded as being at risk for developing the disease because of the mucosal changes typically associated with these conditions. These were gastric ulcer, atrophic gastritis and gastric polyp. A total of 370 cases fell into the high risk group. Class 2 (the low risk group) comprised the remaining cases.

Classification was performed by Predictor. Various aspects of Predictor have been described elsewhere by White & Liu (1993), Liu (1993) and Liu & White (to appear) and will not be described here in detail. Briefly, Predictor operates by a recursive binary partitioning of the data space, under a form of statistical control which branches preferentially on the more important variables. A stopping rule based on significance testing principles (White & Liu, to appear) guards against excessive branching and the specification of a minimum terminal node frequency provides additional control against overfitting. Missing values are dealt with by dynamic path generation, as described in the references just cited. Cross-validation is provided as an option. This is a well-established statistical technique, whose purpose is to provide a fair assessment of the performance of a predictive system. It involves testing each case separately, using a model derived from the other cases.

In order to produce the cross-validated sensitivity-specificity curve, the significance level in Predictor was set at 0.5 and the minimum terminal node frequency at 5. Cross-validation mode was used. Another option in Predictor produces the posterior probabilities of class membership for each case classified. This feature was employed and the resulting probabilities were post-processed by separate software to produce the data required for the sensitivity-specificity graph.

4 Results and Discussion

Part of the cross-validated sensitivity-specificity curve is displayed in Figure 1. It should be noted that a large proportion of spurious missing values was discovered on some of the variables, due to problems encountered in the data before they were transferred to the computer used for running Predictor. Nevertheless, the machine learning algorithm was able to perform competently in spite of this difficulty.

Some idea of the power of the discrimination may be gained by looking closely at the graph in Figure 1. Most of the curve corresponds to criterion settings giving odds ratios (for the cross-validated classification matrix) of 5 or more. Better discrimination is apparent at the high-sensitivity end of the curve, which is beneficial because of the greater interest in performing discrimination with a high-sensitivity criterion because of the seriousness of gastric cancer if diagnosis is missed in the early stages.

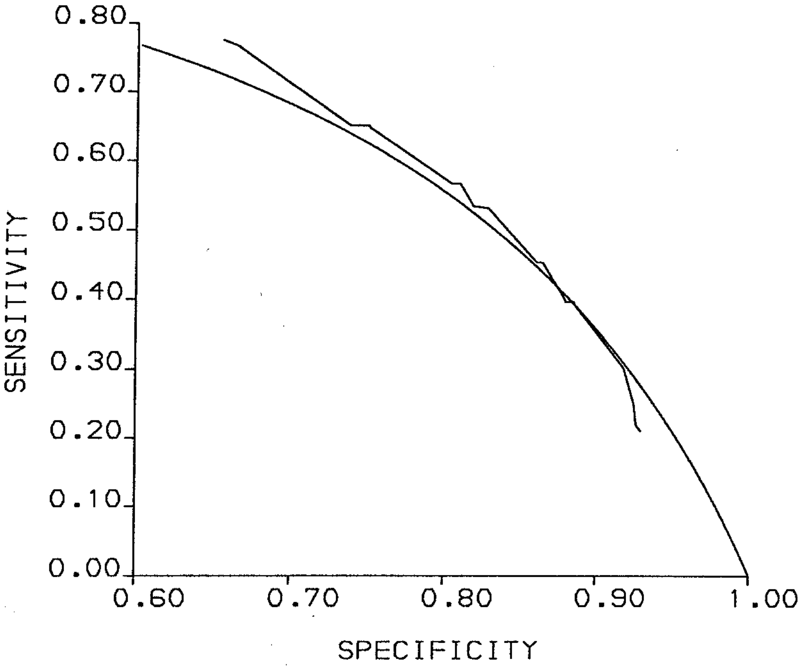

Figure 1: Part of the cross-validated sensitivity-specificity curve for detection of those at high risk for gastric cancer. The smooth curve represents a contour line for an odds ratio of 5. See text for further explanation.

For example, one particular point on this curve (obtained by classifying cases as high-risk if the cross-validated estimated posterior probability of membership of this class is greater than or equal to 0.2) gives a sensitivity of 0.768 and a specificity of 0.664, with an odds ratio of 6.54 for the cross-validated classification matrix. With this criterion setting, 45.0% of the sample are classified as high-risk.

References

Hallissey, M.T., Allum, W.H., Jewkes, A.J., Ellis, D.J. and Fielding, J.W.L. (1990). Early detection of gastric cancer. *British Medical Journal*, **301**, 513-515.

Liu, W.Z. (1993). *Aspects of Machine Learning.* Unpublished PhD thesis.

Liu, W.Z. & White, A.P. (to appear). The importance of attribute selection measures in decision tree induction. *Machine Learning.*

White, A.P. and Liu, W.Z. (1993). Fairness of attribute selection in probabilistic induction. In *Research and Development in Expert Systems* IX, edited by M.A. Bramer. Cambridge: Cambridge University Press.

White, A.P. and Liu, W.Z. (to appear). Bias in information-based measures in decision tree induction. *Machine Learning.*