Diagnosing Acute Appendicitis with Very Simple Classification Rules

Aleksander Øhrn and Jan Komorowski

Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway

Abstract. A medical database with 257 patients thought to have acute appendicitis has been analyzed. Binary classifiers composed of very simple univariate if-then classification rules (1R rules) were synthesized, and are shown to perform well for determining the true disease status. Discriminatory performance was measured by the area under the receiver operating characteristic (ROC) curve. Although an 1R classifier seemingly performs slightly better than a team of experienced physicians when only readily available clinical variables are employed, an analysis of cross-validated simulations shows that this perceived improvement is not statistically significant ($p < 0.613$). However, further addition of biochemical test results to the model yields an 1R classifier that is significantly better than both the physicians ($p < 0.03$) and an 1R classifier based on clinical variables only ($p < 0.0003$).

1 Introduction

Acute appendicitis is one of the most common problems in clinical surgery in the western world, and its diagnosis is sometimes difficult to make, even for experienced physicians. The costs of the two types of diagnostic errors in the binary decision-making process are also very different. Clearly, unnecessary operations are desirable to avoid. But failing to operate at an early enough stage may lead to perforation of the appendix. Perforation of the appendix is a serious condition, and leads to morbidity and occasionally death. Therefore, a high rate of unnecessary surgical interventions is usually accepted. Analysis of collected data with the objective of improving various aspects of diagnosis is therefore potentially valuable.

This paper reports on an analysis of a database of patients thought to have acute appendicitis. The main objective of this study has been to address the following two questions: (1) *Based only upon readily available clinical attributes, does a computer model perform better than a team of physicians at diagnosing acute appendicitis?* and (2) *Does a computer model based upon both clinical attributes and biochemical attributes perform better than a model based only upon the clinical attributes?* These two issues have previously been addressed in the medical literature by Hallan et al. [3, 4], using the same database of patients as presently considered. Multivariate logistic regression (MLR), the de facto standard method for analysis of binary data in the health sciences, was used in

J.M. Żytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 462–467, 1999.

those studies. This paper addresses the same issues, but rather using one of the simplest approaches to rule-based classification imaginable, namely a collection of univariate if-then rules. Univariate if-then rules are also referred to as 1R rules.

2 Preliminaries

Let U denote the universe of patients, let A denote the set of classifier input attributes, and let d denote the outcome attribute. The set of 1R rules is defined as all rules on the form "if $(a = a(x))$ then $(d = d(x))$", where $a \in A$ and $x \in U$. 1R rules have previously been investigated by Holte [7].

A binary classifier realizes a composed decision function $\theta_\tau \circ \phi$, where $\phi(x) \in [0, 1]$ measures the classifier's certainty that a patient x has outcome 1. The function θ_τ is a simple threshold function that evaluates to 0 if $\phi(x) < \tau$, and 1 otherwise. By varying τ and plotting the resulting true positive rates against the false positive rates, one obtains a receiver operating characteristic (ROC) curve. ROC analysis is a graphical method for assessing the discriminatory performance of a binary classifier [5], independent of both error costs and the prevalence of disease. The area under the ROC curve (AUC) is of particular interest, as it equals the Wilcoxon-Mann-Whitney statistic. An AUC of 0.5 signifies that the classifier performs no better than tossing a coin, while an area of 1.0 signifies perfect discrimination.

3 Methodology

For employing 1R rules, discretization of numerical attributes is a necessary prerequisite. In this study, for simplicity, all numerical attributes were discretized using an equal frequency binning technique with three bins, intuitively corresponding to labeling the values "low", "medium" or "high" relative to the observations.

To make the most out of scarce data, k-fold cross-validation (CV) was employed. In the training stage of the CV pipeline, the union of the $k - 1$ blocks were first discretized. 1R rules were subsequently computed from the discretized union of blocks. In the testing stage, the hold-out block was first discretized using the same bins that were computed in the training stage, and the cases in the discretized hold-out block were then classified using standard voting among the previously computed 1R rules.

The results from the voting processes among the 1R rules were used to construct ROC curves. Performance measures for each iteration were harvested by computing the area under the ROC curves (AUC), computed using the trapezoidal rule for integration, as well as their associated standard errors as determined by the Hanley-McNeil formula [5].

Two variations of k-fold CV were applied. First, a single 10-fold CV replication was performed, corresponding to how CV is traditionally employed.

Additionally, five different replications of 2-fold CV was performed, as proposed by Alpaydin [1].

The outlined procedure was done for the three different classifiers below, with identical divisions into k blocks across all classifiers.

- *Simple 1R*: A 1R computer model, based only upon readily available clinical attributes.
- *Extended 1R*: A 1R computer model, based upon the same attributes as the simple 1R model, but with additional access to the results of certain biochemical tests.
- *Physicians*: A classifier realized by probability estimates given by a team of physicians, based upon the same attributes as the simple 1R classifier.

Lastly, a statistical analysis comparing their differences was performed using the methods of Hanley and McNeil [6] and Alpaydin [1].

4 Experiments

The methodology outlined in Sec. 3 has been applied to a medical database with 257 patients thought to have acute appendicitis, summarized in Tab. 1. The 257 patients were referred by general practitioners to the department of surgery at a district general hospital in Norway, and were all suspected to have acute appendicitis after an initial examination in the emergency room. Attributes $\{a_1, \ldots, a_{14}\}$ are readily available clinical attributes, while attributes $\{a_{15}, \ldots, a_{18}\}$ are the results of biochemical tests. The outcome attribute d is the final diagnosis of acute appendicitis, and was based on histological examination of the excised appendix.

After the clinical variables were recorded the physician also gave an estimate of the probability that the patient had acute appendicitis, based on these. Nine residents with two to six years of surgical training participated in the study. These estimates directly define a realization of the certainty function ϕ.

For a detailed description of the patient group and the attribute semantics, see [3, 4].

5 Results

The mean AUC scores from the 10-fold CV simulation with mean standard errors in parentheses were 0.823 (0.089) for the physicians, 0.858 (0.083) for the simple 1R classifier, and 0.920 (0.060) for the extended 1R classifier. The same scores from the five 2-fold CV simulations were 0.818 (0.041), 0.838 (0.039) and 0.910 (0.030), respectively. All simulations were carried out using the ROSETTA software system [8]. On average, the extended 1R classifier seemed to perform somewhat better than both the simple 1R classifier and the team of physicians. The simple 1R classifier and the physicians seemingly perform approximately the same, with the former achieving a slightly better average score.

Attribute		Description	Statistics
a_1	*AGE*	Age (years)	3–86 (22)
a_2	*SEX*	Male sex?	55.3%
a_3	*DURATION*	Duration of pain (hours)	2-600 (22)
a_4	*ANOREXIA*	Anorexia?	69.3%
a_5	*NAUSEA*	Nausea or vomiting?	70.8%
a_6	*PREVIOUS*	Previous surgery?	9.3%
a_7	*MOVEMENT*	Aggravation of pain by movement?	61.5%
a_8	*COUGHING*	Aggravation of pain by coughing?	59.9%
a_9	*MICTUR*	Normal micturation?	87.2%
a_{10}	*TENDRLQ*	Tenderness in right lower quadrant?	86.0%
a_{11}	*REBTEND*	Rebound tenderness in right lower quadrant?	55.3%
a_{12}	*GUARD*	Guarding or rigidity?	30.7%
a_{13}	*CLASSIC*	Classic migration of pain?	49.4%
a_{14}	*TEMP*	Rectal temperature (°C)	36.4–40.3 (37.7)
a_{15}	*ESR*	Erythrocyte sedimentation rate (mm)	1–90 (10)
a_{16}	*CRP*	C-reactive protein concentration (mg/l)	0–260 (12)
a_{17}	*WBC*	White blood cell count ($\times 10^9$)	2.9–31 (12.1)
a_{18}	*NEUTRO*	Neutrophil count (%)	38–93 (80)
d	*DIAGNOSIS*	Acute appendicitis?	38.1%

Table 1: Summary of attributes recorded for the 257 patients thought to have acute appendicitis. For binary attributes, the prevalence is given. For numerical attributes, the range and median are given.

It is trivial to produce a classifier that classifies the training data perfectly. Although this would be a very optimistically biased estimate, 1R rules are so simple they do not possess enough degrees of freedom to overfit the data much. Reference ROC curves obtained when applying the classifiers to the full set of 257 patients from which they were constructed are displayed in Fig. 1.

The exact same set of 257 patients has been previously analyzed by Hallan et al. [3] using MLR. With a slightly different resampling scheme than the one presently employed, an MLR model based upon only the clinical attributes had a mean AUC of 0.854, while an MLR model based on both the clinical attributes and the biochemical attributes had a mean AUC of 0.920. Carlin et al. [2] have also analyzed the same set of patients, but used rough set (RS) methods. Using a similar resampling scheme as Hallan et al., the same scores were 0.850 and 0.923, respectively[1].

6 Analysis

In order to draw any trustworthy conclusions from the results in Sec. 5, a statistical analysis has been performed. The standard tool for comparing correlated AUC values is Hanley-McNeil's method [6]. However, this method is usually employed for a single two-way split only and not in a CV setting. The five 2-fold CV results have been analyzed using the method of Hanley and McNeil on a per-fold per-replication basis. Considering the median p-values, there is no significant difference between the physicians and the simple 1R classifier ($p < 0.585$). On

[1] Both Hallan et al. and Carlin et al. included slightly fewer attributes. However, this was done because Hallan et al. found that adding more attributes did not improve the models further.

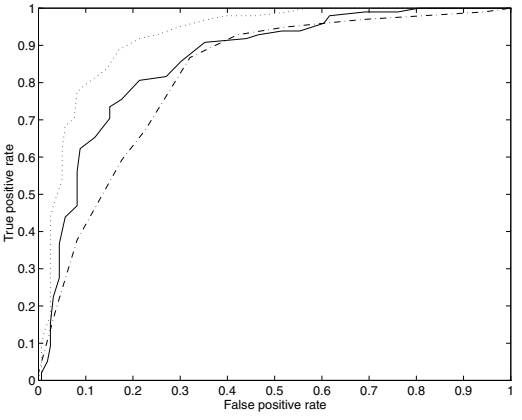

Fig. 1: Reference ROC curves. The middle solid line represents the simple 1R classifier, while the top dotted line represents the extended 1R classifier. The physicians are represented by the bottom dashed line. The AUC values and their standard errors of the three classifiers are 0.817 (0.029) for the physicians, 0.859 (0.026) for the simple 1R classifier, and 0.924 (0.019) for the extended 1R classifier.

the other hand, the extended 1R classifier is significantly better than both the physicians ($p < 0.026$) and the simple 1R classifier ($p < 0.018$).

Simply averaging the p-values as done above to obtain a summary p-value does not capture any systematic variations in differences in performance across folds and replications, information which is obviously of importance. There are, however, statistical tests that have been specifically designed for combining CV together with detection of differences in performance. Applying the 5x2CV F-test of Alpaydin [1] to the five 2-fold CV results again yields similar conclusions. There is no significant difference between the physicians and the simple 1R classifier ($p < 0.613$), but the extended 1R classifier is significantly better than both the physicians ($p < 0.03$) and the simple 1R classifier ($p < 0.0003$).

7 Discussion

In Sec. 1, it was argued that performing a large number of unnecessary operations was preferable to missing any cases of acute appendicitis. This corresponds to prioritizing test sensitivity before test specificity. As can be seen from Fig. 1, the simple 1R classifier and the physicians display virtually identical performance in the area of ROC space of interest, while the extended 1R classifier outperforms them both everywhere.

Simulations by Holte [7] showed that the best individual 1R rules were usually able to come within a few percentage points of the error rate that more complex models can achieve, on a spread of common benchmark domains. The present study suggests that this might be true for other performance measures, too.

8 Conclusions

Based on the results in Sec. 5 and the analysis in Sec. 6, the answers to the two main questions raised in Sec. 1 are: (1) *No, not significantly, at least not with a set of very simple 1R classification rules as the computer model*, and (2) *Yes, even with a set of very simple 1R classification rules as the computer model there is a significant improvement when biochemical attributes are additionally taken into account.*

It hardly seems likely that the almost identical results reported in the literature and repeated in Sec. 5 based on MLR [3, 4] or complex RS models [2] are statistically significantly different from the 1R results reported in this study. Hence, based on the principle of parsimony, a collection of very simple 1R classification rules seems like a good rule-based candidate for diagnosing acute appendicitis as measured by the area under the ROC curve.

Acknowledgments

Thanks to Stein Hallan and Arne Åsberg for sharing the appendicitis data, and to Tor-Kristian Jenssen and Ulf Carlin. This work was supported in part by grant 74467/410 from the Norwegian Research Council.

References

[1] E. Alpaydin. Combined 5x2CV F test for comparing supervised classification learning algorithms. Research Report 98-04, IDIAP, Martigny, Switzerland, May 1998. To appear in Neural Computation.

[2] U. Carlin, J. Komorowski, and A. Øhrn. Rough set analysis of patients with suspected acute appendicitis. In *Proc. Seventh Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'98)*, pages 1528–1533, Paris, France, July 1998. EDK Éditions Médicales et Scientifiques.

[3] S. Hallan, A. Åsberg, and T.-H. Edna. Additional value of biochemical tests in suspected acute appendicitis. *European Journal of Surgery*, 163(7):533–538, July 1997.

[4] S. Hallan, A. Åsberg, and T.-H. Edna. Estimating the probability of acute appendicitis using clinical criteria of a structured record sheet: The physician against the computer. *European Journal of Surgery*, 163(6):427–432, June 1997.

[5] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, Apr. 1982.

[6] J. A. Hanley and B. J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843, Sept. 1983.

[7] R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1):63–91, Apr. 1993.

[8] A. Øhrn, J. Komorowski, A. Skowron, and P. Synak. Rough sets in knowledge discovery 1: Methodology and applications. volume 18 of *Studies in Fuzziness and Soft Computing*, chapter 19, pages 376–399. Physica-Verlag, Heidelberg, Germany, 1998.