ANALYSIS OF OVARIAN TUMOR DATA: EXTENDED SUMMARY

Dyre Tjeldvoll, Jan Komorowski
Knowledge Systems Group, Department of Information and Computer Science
Norwegian University of Science and Technology
7491 Trondheim, Norway
Phone: +47 73593440, Fax: +47 73594466
e-mail: {janko, dyret}@idi.ntnu.no

Lil Valentin
Department of Obstetrics and Gynaecology
University of (Lund / Malmö)
Löviksg 7B, S21374 Malmö, Sweden
Phone: +46 40 949726, Fax: +46 40 949726
e-mail: lil.valentin@obst.mas.lu.se

Dirk Timmermann
Department of Obstetrics and Gynaecology
University Hospitals Leuven,
Herestraat 49, B-3000 Leuven, Belgium
Phone: +32 16 344215, Fax: +32 16 344205
e-mail: dirk.timmerman@uz.kuleuven.ac.be

ABSTRACT: The goal was to use rough sets and the software package rosetta to develop a classifier and a decision protocol, for use in diagnosing ovarian cancer. The data set is made up of 300 objects with about 20 attributes, including one binary decision attribute indicating the presence or absence of malignancy.

KEYWORDS: Medical diagnosis, rough sets, rosetta, ovarian cancer

INTRODUCTION

Malignant ovarian tumors result in the highest mortality figures of all gynecological cancers, so the ability to diagnose such tumors correctly would be very valuable indeed. Furthermore, it is very beneficial both for the patients chances of survival, and for medical costs involved, if a reliable diagnosis can be obtained prior to surgery [1].

The problem of obtaining such a diagnosis is difficult. However, experienced ultrasonographers can make very good classifications based on the ultrasonographic image, and (subconsciously) on additional clinical information which they may have. Typically they achieve a sensitivity of about 90% and a specificity of 96%, but it is worth noting that this is based on experience from more than 4000 examinations, and an ultrasonographer with experience from more than 10000 examinations did not perform significantly better [1].

With the use of rough sets and rosetta [2, 3, 4] one would like to develop a classifier that can perform as well, or better than an ultrasonographer, and if that is not possible; to develop a classifier that can aid less experienced ultrasonographers. Several other methods have been employed in attempts to accomplish this, including logical regression analysis and artificial neural networks, but so far they have not been able to perform as well as a human examiner.

The rough set approach makes it possible to formalize knowledge (in the form of rules), which currently only the experienced examiners have. Hence, if the number of rules can be kept at reasonable level, this knowledge can be distributed to

less experienced examiners.

METHODOLOGY

Rosetta is a rough set analysis toolkit developed both at NTNU and the University of Warsaw. Rosetta is capable of computing reducts, that is minimal sets of attributes capable of discerning between different decision classes, and to induce rules based on these reducts. These rules can then be classified through applying them to a different, not yet seen, part of the data set. The results of the classification are accuracy for the chosen threshold value, usually 0.5, a trapezoidal approximation of the area under the ROC curve (AUC), the threshold value closest to the point (0,1) on the ROC curve (Thr(0,1)), and the threshold value resulting in the highest accuracy (Thr.acc.). Rosetta also calculates the data points needed to plot the ROC curves. A (meaningful) reduct computation can only be performed on discrete data sets. Consequently Rosetta is equipped with various discretization algorithms that can be applied prior to reduct computation.

Rosetta was used to perform rough set analysis on data obtained from the examination of 300 women suspected of having ovarian cancer. The original data set was split into a training set of 191 objects and a test set made up of the remaining 109 objects. All the analysis that have been carried out so far, have been made on the training set, but the evaluation of the final model will be made on the test set. For each trial the original training set is split into a new training set containing 91 objects, and a new test set containing 46 objects. The 54 objects not accounted for were dropped because they had one or more undefined (missing) attribute values.

Originally the data set had 41 attributes, but only about 20 of these would be used by an expert when predicting an outcome. This simplification made the analysis considerably more manageable. The resulting data set then has one binary decision attribute, *"Pathology"*, which states whether the tumor in question is benign or malignant. Of the remaining attributes, 8 were continuous, 4 were multi-valued discrete, and the rest were binary discrete. The continuous attributes needed be discretized before rough set analysis could be performed. The chosen solution was to use manual discretization with cut-off values supplied by an expert, in the form of values suggested in articles presenting results from previous studies [1].

Rules were induced on the discretized data set by applying the Genetic Reducer Algorithm in rosetta. On a data set with this number of attributes the algorithm takes a significant amount of time to complete, and produces a rather large rule set. Consequently, attempts were made at using the much faster Johnson Reducer Algorithm which also produces much fewer rules. Unfortunately the resulting rules failed to classify several objects in test set. As a result the use of the Johnson Algorithm was abandoned.

RESULTS

The performance of a classifier is obviously a function of the split between the training and test sets. Hence it is important to test the performance of the classifier using different seeds for the random number generator employed by the splitting algorithm, when making the splits. This has been done, both by comparing the results obtained using different seeds directly and also by performing cross-validation. This will ensure that the results are not due to extremely favorable or unfavorable splits.

The classifiers producing the results presented here are the result of ongoing research. Hence, the results are expected to improve as the model becomes more refined and adapted to the data set. Some methods that will be employed to achieve this, are presented later in this paper. The models for which results are presented here, have been developed by splitting the original training set in two, using the Binary Splitter Algorithm provided by rosetta. It is possible to adjust the ratio between the size of the two resulting sets. For the results presented here a ratio of 0.667 was used, giving one set with 91 objects and one with 46 objects, where the largest set was used as the new training set. The splits were made with six different seeds; 6,7,8,9,10,11, and in each case a model was obtained by computing reducts with the SAVGeneticReducer function, and inducing rules with the RSESRuleGenerator function. The results are summarized in the table below:

RNG Seed	AUC	Accuracy	Thr(0,1)	Thr.acc.
6	0.961905	0.804348	0.136	0.136
7	0.984127	0.913043	0.252	0.32
8	0.936508	0.891304	0.148	0.512
9	0.928571	0.934783	0.3	0.556
10	0.965625	0.869565	0.128	0.128
11	0.941176	0.934783	0.308	0.308
Average	0.952985	0.891304	0.212	0.327

For cross-validation, a 5-fold cross-validation was performed using the CVSerialExecutor functionality in rosetta, again using the SAVGeneticReducer to compute reducts, and the RSESRuleGenerator to induce rules. From the cross-validation the following average values were computed:

AUC	Accuracy	Thr(0,1)	Thr.acc.
0.955912	0.904725	0.2592	0.352

From this it would appear that a lower threshold would result in higher accuracy. Applying cross-validation with a classifier using a threshold of 0.35, the following results were obtained:

AUC	Accuracy	Thr(0,1)	Thr.acc.
0.959103	0.912133	0.2432	0.3384

This is not a very substantial increase in accuracy, especially when taking into account that the AUC value also is slightly higher. A comparison of the two values using statistical methods is not available at present. The fact that changing the threshold to a value close to the average of the maximizing thresholds from the cross-validation did not drastically increase the average accuracy, can be attributed to the relatively high variation in the maximizing thresholds.

ANALYSIS

At first sight the data set seemed ideally suited for rough set analysis. But, as is often the case with real-world data, there were problems that had to be resolved before the analysis could be performed. One such difficulty concerned the internal dependencies in the dataset. Most striking of these were link between the attribute *"Menopausal status"* and *"Years since menopause"* and *"Cycle day"*. For pre-menopausal objects the *"Years since menopause"* attribute was, of course, undefined. Likewise for post-menopausal objects the *"Cycle day"* attribute was undefined. This led to a situation where every object had one attribute with an undefined value. Having undefined values in the data set is highly undesirable, but finding satisfactory solution proved to be difficult.

One possible solution would be to collapse the three attributes into one where a positive number indicates *"Cycle day"*, a negative number indicates *"Years since menopause"* (or vice versa) and zero indicates a hysterectomized object. But this would turn an important discrete attribute (*"Menopausal status"*) into a continuous variable which would be difficult to discretize. The other solution would be to eliminate the two attributes altogether. This meant that potentially important information would be lost. When confronted with the problem, the medical expert suggested that the attributes be dropped because it was doubtful that they were really significant.

Another problem was the large number of missing attribute values. Rosetta provides algorithms that attempts to fill in such missing values. However, these algorithms were not suitable in this particular case, because the probable value for the missing attributes depends on one, or more, of the other attributes in addition to the decision attribute. Rosetta currently only supports fill conditioned on the decision attribute. (In practice on any attribute, since it can be swapped with the decision attribute when filling, and then back again when the filling is completed). Since 54 of 191 objects were incomplete, dropping them represented a severe reduction in the size of the dataset. But because it was so difficult to find a viable method for filling in the missing values, it was deemed to be the best option, at least until a better completion method can be found.

As mentioned previously, the data set contained continuous attributes that required discretization. Unfortunately the distribution of values for the continuous attributes in the two decision classes overlapped considerably. The result was that the discretization algorithm divided the range of values for the continuous attributes into many small intervals to match the decision. This partition was meaningless because the decision was assumed to be a linear function of the particular attribute.

This assumption is justified by the fact that the field experts, who are able to classify objects with very high accuracy, make the same assumption. Furthermore, the many intervals were so specific that they uniquely identified an object, and hence induced a rule. Needless to say, such rules did not have much predictive power for unseen objects. The fact that the choice for these cut-off values were not determined from the actual data set used, will presumably limit the predictive power of the model. The cut-off values currently used to discretize the data set have mostly been taken form previous experiments, based on different datasets. Hence, discretizations that produce better results probably exist. The field expert has been asked to compile a list of presumed optimal cut-off values for the relevant attributes, but this list is not presently available.

DISCUSSION

The only indication of the quality of the model developed so far, are the AUC and accuracy figures presented previously. However, these numbers do not mean much by themselves. Ultimately, one would like to compare this model to those obtained using different methods. So far, no other methods have been applied to this data set, but in [1] logistic regression analysis and artificial neural networks (ANNs) were applied to a comparable data set. A comparison of those results with the results obtained with rough sets and rosetta is presented below (AUC values):

Rosetta	Log. Reg.	ANN 1	ANN 2
0.955912	0.904	0.951	0.967

Although a comparison using statistical methods has not yet been carried out, it seems unlikely that there is any significant difference, given that [1] reports that there is no significant difference between logistic regression and any of the ANNs, and the number of objects used in the two studies is similar. Although the results from ANN 2 seem impressive, [1] reports that ANN 2 performed significantly worse than a human examiner when confronted with a new and slightly different data set. It would seem reasonable to expect that the preliminary rough set model would have similar problems, although no evidence exists that can corroborate or refute this. The data set used to generate the rough set model, contains information about how a group of medical experts classified the objects. Using this data it will be possible to estimate the performance of the model, compared to that of a human examiner.

However, even if the final model should fail to perform significantly better than a human expert or even an AAN, the rough set approach still has the advantage of being able to offer insight into how it makes its predictions. Hence it will be possible to suggest a decision protocol that can be used by less experienced examiners as well as in software. In order to develop a decision protocol that is useful, it will be necessary to prune the rule sets. Quite a lot of work remains before it is possible to determine which rules can be pruned without significantly reducing the quality of the classification [5].

The rosetta System has a dictionary facility which can be used to produce rules that are easier for humans to interpret. Since the data set is, almost exclusively, made up of numerical values, such a dictionary will have to be made before the decision protocol can be reviewed by a human expert. In addition the results will need to be validated using conventional statistical methods.

FUTURE RESEARCH

The number of objects used to generate the model is rather small (91), and this obviously limits its predictive power. It would be interesting to see if a larger data set will improve the results, and perhaps make the classifier more robust. This issue is also discussed in [1], where it is suggested that the large number of different medical conditions that fall in the category malignant tumor, warrants a large data set to make sure that all relevant cases are included.

The problem with discretization could also be the subject of further research. Instead of only relying on the medical experts to supply cut-off values, it would also be interesting to develop methods that can split the objects into a predetermined number of intervals, based on the decision attribute. This functionality is currently not present in rosetta, so additional software will have to be used. The constructed cut-offs could then be compared to those suggested by the experts.

If the decision protocol proved to be successful, it could be used as the basis for an expert system which could aid medical personell when diagnosing tumors. This task would be greatly simplified by the fact that rosetta has the possibility to export its rules to other formats, among them prolog clauses.

ACKNOWLEDGMENTS

Many thanks to Aleksander Øhrn for his many helpful suggestions, and for taking the time to answer our countless questions.

REFERENCES

1. *Ultrasonography in the assessment of ovarian and tamoxifen-associated endometrial pathology*, (D. Timmermann), Leuven University Press 1997.
2. *Rosetta and Rough Set Software Systems for Data Mining and Knowledge Discovery*, (J. Komorowski, A. Skowron and A. Øhrn), to appear in *Handbook of Data Mining and Knowledge Discovery*, (W. Klösgen, J. Zytkow, Eds.), Oxford University Press, 1999.
3. *The Design and Implementation of a Knowledge Discovery Toolkit Based on Rough Sets - The Rosetta System*, (A. Øhrn, J. Komorowski, A. Skowron, P. Synak), in: *Rough Sets in Knowledge Discovery*, (A. Skowron, L. Polkowski, Eds.), Springer Verlag, 1998.
4. *ROSETTA, Technical Reference Manual*, (A. Øhrn) http://www.idi.ntnu.no/∼aleks/rosetta
5. *Finding Small High Performance Subsets of Induced Rule Sets: Extended Summary*, (T. Aagotnes) Submitted for publication, EUFIT 1999.