A Breast Cancer Diagnosis System: A Combined Approach Using Rough Sets and Probabilistic Neural Networks

Kenneth Revett, Florin Gorunescu, Marina Gorunescu, Elia El-Darzi and Marius Ene

Abstract — **In this paper, we present a medical decision support system based on a hybrid approach utilising rough sets and a probabilistic neural network. We utilised the ability of rough sets to perform dimensionality reduction to eliminate redundant attributes from a biomedical dataset. We then utilised a probabilistic neural network to perform supervised classification. Our results indicate that rough sets was able to reduce the number of attributes in the dataset by 67% without sacrificing classification accuracy. Our classification accuracy results yielded results on the order of 93%.**

Keywords — **breast cancer diagnosis, dimensionality reduction, medical decision support systems, Probabilistic Neural Networks, and rough sets**

I. Introduction

BREAST cancer is the most common form of cancer in women in the US and Europe. Diagnosis is generally provided by a mammogram, with additional support provided by cytological examination, and the experience and opinion of the attending surgeon. In the case of a positive diagnosis, in many cases a radical mastectomy is to be performed. It is critical that the accuracy of the diagnosis is 100% - or as close as humanly possible because the treatment for the disease can be permanent and drastic (i.e. radical mastectomy). Therefore, most medical institutions will insist that their personnel err on the side of caution minimising the risk by maximising the specificity of the diagnosis. Generally this results in a reduction in sensitivity as many patients with symptoms falling on the tails of the distribution will not be properly diagnosed. Sensitivity reflects the level of false negatives, which must be below acceptable levels on the order of less than 1% is desirable. How this result is to be achieved is still uncertain. In this paper, we study a well structured and complete breast cancer dataset that has been used as a bench mark test for various machine learning techniques [1]. We are proposing a hybrid decision support system, combining the reductionist approach of rough sets in combination with a probabilistic neural network. Rough sets have been used in various medical diagnostic systems with a large degree of success [2,3]. One of the hallmark features of rough sets is the ability to remove redundant attributes [4]. With the dataset reduced to essential attributes, we then apply a probabilistic neural network (PNN) for the final classification task. The result of this hybrid approach is an extremely accurate classifier (with respect to sensitivity and specificity) that is also computationally efficient . This paper is organised as follows: in the next section we present a brief description of the rough set and the PNN algorithms, followed by a description of the dataset, then a results section followed by a conclusion and future work.

II. Rough Sets

Our hepatic cancer diagnosis pre-processing/classifier is based on the concept of approximate reducts derived from the data-mining paradigm of the theory of Rough Sets [4],[5]. We divide the table into training and test cases, employing N-fold cross validation. The data set is transformed into a decision table (DT) from which rules are generated to provide an automated classification capacity. In generating the decision table, each row consists of an observation (also called an object) and each column is an attribute, with the last one as the decision for this object {d}. Formally, a DT is a pair A = (*U*, A∪{d}) where d A is the *decision attribute,* where *U* is a finite non-empty set of objects called the *universe* and A is a finite non-empty set of attributes such that a:U->V_a is called the value set of a. Rough sets seeks data reduction through the concept of equivalence classes (through the indiscernibiliy relation). By generating such classes, one can reduce the number of attributes in the decision table by selecting any member of the equivalence class as a representation of the entire class. This process generates a series of *reducts* which are subsequently used in the classification process. Finding the reducts is an NP-hard problem, but fortunately there are good heuristics that can

K Revett is with the Harrow School of Computer Science, University of Westminster, London, UK (phone: 44-2079115000; fax: 44-2079115609; e-mail: revettk@westminster.ac.uk).

F. Gorunescu, is with the Department of Mathematics, Biostatistics and Computer Science, University of Medicine and Pharmacy of Craiova, Romania (e-mail: fgorun@rdslink.ro); *IEEE member*

M. Gorunescu is with the Department of Computer Science, Faculty of Mathematics and Computer Science, University of Craiova, Romania (e-mail: mgorun@inf.ucv.ro)

E. El-Darzi is with the Harrow School of Computer Science, University of Westminster, London, UK (e-mail: eldarze@westminster.ac.uk); *IEEE member*

M. Ene is with the Department of Mathematics, Biostatistics and Computer Science, University of Medicine and Pharmacy of Craiova, Romania (e-mail: enem@umfcv.ro)

1-4244-0049-X/05/$20.00 ©2005 IEEE

compute a sufficient amount of approximate reducts in reasonable time to be usable. In the software system that we employ (developed by 1 of the authors) an order based genetic algorithm (o-GA) ([6]) is used to search through the decision table for approximate reducts which result in a series of 'if..then..' decision rules. We then apply these decision rules to the test data and measure specificity and sensitivity of the resulting classifications. In addition, we examined in a systematic fashion, which attributes were most informative in the decision process this can be determined by examining the correlation, coverage, and support of the attributes in the final set of decision rules. This provides us with the entry point for using the probabilistic neural network approach, which we describe in the next section.

III. PROBABILISTIC NEURAL NETWORKS

The PNNs are basically classifiers. The general classification problem is to determine the category membership of a multivariate sample data (i.e. a p-dimensional random vector $\mathbf{x}$) into one of q possible groups Ω_i, $i = 1, 2,\dots, q$, based on a set of measurements. If we know the probability density functions (p.d.f.) $f_i(\mathbf{x})$, usually the Parzen-Cacoulos or Parzen like p.d.f. classifiers:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}\sigma^p}\cdot\frac{1}{m_i}\cdot\sum_{j\ 1}^{m_i}\exp\left(-\frac{\left\|x-x_j\right\|^2}{2\sigma^2}\right), \qquad (1)$$

the *a priori* probabilities $h_i = P(\Omega_i)$ of occurrence of patterns from categories Ω_i and the *loss* (or *cost*) parameters l_i associated with all incorrect decisions given $\Omega = \Omega_i$, then, according to the Bayesian decision rule, we classify $\mathbf{x}$ into the category Ω_i if the inequality $l_i\ h_i\ f_i(\mathbf{x}) > l_j\ h_j\ f_j(\mathbf{x})$ holds true. The standard training procedure for PNN requires a single pass over all the training patterns, giving them the advantage of being faster than the feed-forward neural networks [7].

Basically, the architecture of PNN is limited to three layers: the *input/pattern layer*, the *summation layer* and the *output layer*. Each input/pattern node forms a product of the input pattern vector $\mathbf{x}$ with a weight vector W_i and then perform a nonlinear operation, that is $\exp[-(W_i - x)^\tau (W_i - x)/(2\sigma^2)]$ (assuming that both $\mathbf{x}$ and W_i are normalized to unit length), before outputting its activation level to the summation node. Each summation node receives the outputs from the input/pattern nodes associated with a given class and simply sums the inputs from the pattern units that correspond to the category from which the training pattern was selected, $\sum_i \exp[-(W_i - x)^\tau (W_i - x)/(2\sigma^2)]$. The output nodes produce binary outputs by using the inequality:

$$\sum_i \exp[-(W_i - x)^\tau (W_i - x)/(2\sigma^2)] > $$
$$\sum_j \exp[-(W_j - x)^\tau (W_j - x)/(2\sigma^2)], \qquad (2)$$

related to two different categories Ω_i and Ω_j. The key to obtain a good classification using PNN is to optimally estimate the two parameters of the Bayes decision rule, the misclassification costs and the prior probabilities. In our practical experiment we have estimate them heuristically. Thus, as concerns the costs parameters, we have considered them depending on the average distances D_i, inversely proportional, that is $l_i = 1/D_i$. As concerns the prior probabilities, they measure the membership probability in each group and, thus, we have considered them equal to each group size, that is $h_i = m_i$. As in our previous work, we employed an evolutionary technique based on the genetic algorithm to find the smoothing parameters (cf [8] for implementation details). To avoid overfitting, the data set was randomly partitioned into two sets: the training set and the validation set. A number of 458 persons (70%) of the initial group were withheld from the initial group for the smoothing factor adjustment (the training process). Once optimal smoothing parameters σ's for each decision category were obtained using the training set, the trained PNN was applied to the validation set (the remaining 241 persons).

IV. DECISION TABLE DESCRIPTION

We utilised the Wisconsin Breast cancer dataset which contains nine numeric discrete attributes for a set of 699 patients, (16 missing values) [1]. The attribute labels and their corresponding value ranges are listed in Table 1 below. For the missing values, we employed conditioned median imputation for the rough sets algorithm. For the PNN algorithm, we proceeded in two ways: i) we omitted entries in the table with missing values (yielding 683 entries) or used the conditioned median imputation from the rough sets algorithm, yielding a complete table with all 699 entries. There were 458 benign cases (65.6%) and 241 malignant cases (34.4%). We divide the table into training and test cases, using a 50/50 split respectively. employing 10-fold cross validation (350 training/349 testing). We partitioned the dataset randomly with replacement 10 times, each time selecting out 50% for training and 50% for testing purposes. The first goal of this work entailed reducing the size of the data by eliminating any non-informative attributes. We determined the Pearson correlation coefficients of the attributes with respect to their decision class as a first estimation of which attributes could be removed from the decision table. The results are presented in Table 2.

V. RESULTS

Table I present a description of the attribute labels and the value ranges for all nine attributes in the Wisconsin Breast cancer dataset. The three bold entries were the attributes of the reduced dataset as discussed in the text. The data in the decision table was completely discretised, and we imputed missing values (16 of them) using a conditioned median filling method available within Rosetta, an implementation of rough sets available from the internet [9]. The decision table was split 70/30 training and

TABLE 1: Wisconsin Breast cancer Dataset description

Patient Number	NOT USED
Clump Thickness	1-10
Uniformity of Cell Size	1-10
Cell Shape Uniformity	1-10
Marginal Adhesion	1-10
Single- Epithelial Cell Size	1-10
Bare-Nuclei	1-10
Bland-Chromatin	1-10
Normal-Nucleoli	1-10
Mitoses	1-10
Decision Class	**0 = Benign,** **1 = Malignant**

TABLE 2: Pearson Correlation coefficients for all attributes used in the decision table. Attributes with a '*' indicate highest correlation coefficient

Clump Thickness	**0.71**
Uniformity of Cell Size	**0.82 ***
Uniformity of Cell Shape	**0.82 ***
Marginal Adhesion	**0.69**
Single- Epithelial Cell Size	**0.68**
Bare-Nuclei	**0.80 ***
Bland-Chromatin	**0.75**
Normal-Nucleoli	**0.70**
Mitoses	**0.42**

testing with 10-fold cross validation. We generated approximate reducts using the exhaustive RSES facility. From the collection of approximate reducts, we generated the decision rules that were used to classify all objects in the decision table to their respective decision class ('0' = benign or '1' = malignant'). In addition, based on the correlation coefficients (depicted in Table 2), we masked off attributes from the decision table that had a correlation coefficient below a given threshold. The 5-attribute threshold was 0.70, and the 3-attribute threshold was 0.79. In table 3, we present an example classification from the confusion matrix with all 9, 5 and 3 attributes. The 5-attribute rule generation facility used the attributes with the five largest correlation coefficient and the 3-attribute rules set from the attributes with the top three correlation coefficients. We also tried various exhaustive combinations of attributes to see which provided the best classification accuracy. The results (data not shown) indicate that those highlighted in Table 2 provided the greatest accuracy indicating a positive correlation between the correlation coefficient and class prediction.

The results from this part of the study conclude the contribution of rough sets to this hybrid classifier. The goal of reducing the attribute set has been achieved, and now we report the results from the PNN classifier. Table 3, above, the confusion matrix bold values at the bottom right hand corners of each confusion matrix entry is the overall accuracy, according to the following formula:

$$Acc = TP + TN /(TP + FP + TN + FN) \qquad (3)$$

Lastly, we examined the average number of rules that were generated from each of x-attribute based classification scheme, summarised in Table 4.

TABLE 3: Sample confusion matrices randomly selected for a series of attributes selected based on their correlation coefficients. Please note 'Malig' is short for 'malignant' and 'Attrib' is short for attribute. Please note all calculated values are truncated with rounding to two decimal places.

9 Attrib.	**Benign**	**Malignant**	
Benign	**209**	**23**	0.90
Malig.	**22**	**95**	0.81
	0.90	0.91	**0.87**
5 Attrib.			
Benign	**201**	**31**	0.87
Malig.	**17**	**100**	0.90
	0.92	0.76	**0.86**
3 Attrib.			
Benign	**202**	**36**	0.85
Malig.	**14**	**97**	0.87
	0.94	0.73	**0.86**

TABLE 4:. Summary of the average number of rules (10 trials) for each of the x-attribute classifications

9-attributes	5-attributes	3-attributes
26,453	1,813	81

We split the data into a 70/30 split in order to derive a value for the smoothing function that will be used in the PNN classification algorithm. In addition, we wished to determine whether the 16 missing values would have an impact on the classification accuracy of the PNN. We therefore trained the classifier using two version of the dataset: one where any object with 1 or more missing values were removed and the other with the full set of objects including those with missing attribute values. Initially, we used the full set of attributes in order to compare the results with those obtained from using rough sets. The data for this experiment are presented in Table 5. In addition, we used only the three attributes with the highest correlation coefficient and repeated the procedure as above. w present those results in Table 6. Note that the values reported in Tables 5 & 6 are the average of 10 runs (using 10-fold validation) to enhance the statistical significance of our results.

TABLE 5: Summary of the PNN classification results using the full set of attributes (all 9). The classification results for the training and test cases are presented for the DT without missing values. The 683 columns refer to the decision table with objects removed if they contain missing value(s).

Training		**Testing**	
683	699	683	699
100%	100%	93%	92%

Lastly, we present the results of the PNN classification using only the reduced dataset (i.e. with the attributes with the 3 largest correlation coefficients) in Table 6. We did not test using the partial decision table (i.e. removing objects with missing values).

TABLE 6: Summary of PNN results using the reduced decision table (3-attributes the same ones in the last entry in Table 3)

Training	**Testing**
99%	**85%**

IV. CONCLUSION

In this study, we present a hybrid classification system incorporating an implementation of rough sets and a probabilistic neural network to classify a well known breast cancer dataset. The rough sets component provided a means of reducing the number of attributes. We used the Pearson correlation coefficient as a starting point for reducing the number of attributes, based on a user defined threshold. We corroborated that the correlation coefficient was indeed a useful measure of the overall importance of a given attribute with respect to classification accuracy. In this dataset, the correlations were all quite high. But in general, this is not the case. From our experience, many attributes have small Pearson correlation coefficients (many are negative) and therefore one can not rely on this value alone. We tested the effect of removing particular attributes by performing a complete rough sets based classification. We performed this process exhaustively in this particular case. If there are a large number of attributes, then this process can become prohibitive. The PNN algorithm was then employed to perform classification both to see how it performed on this dataset and to have a basis for comparison of the classification results obtained from using rough sets alone. The classification results from the two methodologies were fairly similar to one another, although the PNN generally outperformed the rough set classifier in this instance. The classification results were also consistent with some of the highest results obtained from other classifiers published in the literature [10],[11]. In general, a PNN is usually faster to run, because the pre-processing stages involved in rough sets are not required in a PNN. It can handle missing data items better than rough sets which generally requires imputation. We plan to carry out further experimentation on how these two technologies can be combined in a seamless fashion. The result would produce a classifier that has the advantages of producing a rule based system, suitable for use in an expert-system, along with the noise tolerance of a PNN. This is a typical requirement when working with small biomedical datasets that are filled with missing values.

Acknowledgement: This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

REFERENCES

[1] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[2] K. Revett and A. Khan, "A Rough Sets Based Breast Cancer Decision Support System," METMBS, Las Vegas, Nevada, June, 2005

[3] A. Khan & K. Revett. "Data mining the PIMA Indian diabetes database using Rough Set theory with a special emphasis on rule reduction," INMIC2004, Lahore Pakistan, pp. 334-339, December, 2004

[4] Z. Pawlak . Rough Sets, International Journal of Computer and Information Sciences, 11, pp. 341-356, 1982.

[5] D. Slezak.: "Approximate Entropy Reducts". Fundamenta Informaticae, 2002.

[6] Wroblewski, J.: "Theoretical Foundations of Order-Based Genetic Algorithms". Fundamenta Informaticae 28(3-4) pp. 423 430, 1996.

[7] D. F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, pp. 109-118, 1990.

[8] F. Gorunescu, "Architecture of probabilistic neural networks: estimating the adjustable smoothing factor," Research Notes in Artificial Intelligence and Digital Communications, 104, pp. 56-62, 2004.

[9] Rosetta: Rosetta: http://www.idi.ntnu.no/~aleks/rosetta

[10] P.R. Bakic & D.P. Brzakovic,. "Application of neural networks in computer aided diagnosis of breast cancer," www.eecs.lehigh.edu/~ipal/neurel97_final.ps

[11] M.A. Markey, J.Y. Lo., G.D. Tourassi, , & C.E. Floyd jr. "Self-organzing map for cluster analysis of a breast cancer database," Artificial Intelligence in Medicine, 27, pp. 113-127, 2003.