Attribute Selection for EEG Signal Classification Using Rough Sets and Neural Networks

Kenneth Revett[1], Marcin Szczuka[2],
Pari Jahankhani[1], and Vassilis Kodogiannis[1,3]

[1] Mechatronics Group, Harrow School of Computer Science,
Univ. of Westminster, London HA1 3TP, UK
{parij, revettk}@wmin.ac.uk
[2] Institute of Mathematics, Warsaw University
Banacha 2, 02-097, Warsaw, Poland
szczuka@mimuw.edu.pl
[3] Centre of Systems Analysis and the Mechatronics Group,
School of Computer Science, Univ. of Westminster, London HA1 3TP, UK
kodogiv@wmin.ac.uk

Abstract. This paper describes the application of rough sets and neural network models for classification of electroencephalogram (EEG) signals from two patient classes: normal and epileptic. First, the wavelet transform (WT) was applied to the EEG time series in order to reduce the dimensionality and highlight important features in the data. Statistical measures of the resulting wavelet coefficients were used for the classification task. Employing rough sets, we sought to determine which of the acquired attributes were necessary/informative as predictors of the decision classes. The results indicate that rough sets was able to accurately classify the datasets with an accuracy of almost 100%. The resulting rule sets were small, with an average cardinality of 6. These results were confirmed using standard neural network based classifiers.

1 Introduction

Electroencephalography (EEG) provides a direct measure of cortical activity with millisecond temporal resolution in a non-invasive manner. The technique is widely used in clinical neurophysiological settings and has provided a wealth of diagnostic information for a wide range of neurological deficits [5,7]. Although the underlying technology has not radically changed since its introduction by Hans Berger in 1924, the amount of data that is generated by EEG studies has increased exponentially. Laboratories routinely use 100-electrode arrays and record for more than 24 hours at a sampling rate of over 100 Hz. In response to this wealth of important data generated by EEG studies, many laboratories around the world have developed various techniques for automating the extraction of diagnostically relevant information.

The current trend in EEG analysis employs a multi-stage process: in the first stage, Discrete Wavelet Transforms (DWT) are used as a pre-processing step to decompose the time series into a number of subbands through a process that is

S. Greco et al. (Eds.): RSCTC 2006, LNAI 4259, pp. 408–417, 2006.

essentially a series of low and high pass filters. This pre-processing step effectively performs a dimensionality reduction of the data in preparation for subsequent analysis. In the next phase, a series of attributes are generated from the DWT processed dataset, which is used as input to a classification system. In the present case, we are seeking to produce a classifier that is capable of distinguishing a normal EEG recording from one containing an epileptic seizure segment. This is not a novel task, as many literature reports have presented a variety of techniques. Neural networks and statistical pattern recognition methods have been applied to EEG analysis. Neural Network (NN) detection systems have been proposed by a number of researchers. Pradhan et al. [10] used the raw EEG as an input to a neural network while Weng and Khorasani [11] used an adaptive structure neural network, but his results show a poor false detection rate. Petrosian et al. [9] showed that the ability of specifically designed and trained recurrent neural networks (RNN) combined with wavelet pre-processing, to predict the onset of epileptic seizures both on scalp and intracranial recordings only one-channel of electroencephalogram. In order to provide faster and efficient algorithm, Folkers et al. [4] proposed a versatile signal processing and analysis framework for bioelectrical data and in particular for neural recordings and 128-channel EEG.

In our approach, we are using rough sets to determine which attributes are the most relevant for the classification task. Ningler et al. [8] have applied rough sets to classify EEG-signals with respect to intraoperative awareness, with a reasonable degree of success (90% classification accuracy). They did not however relate these resulting attributes to any underlying phenomenon, as is evidenced by their very large rule set (475 for the crisp and 13,424 for the fuzzy discretisation methods respectively). We applied rough sets to a set of attributes (measures of dispersion from the DWT pre-processing step) in order to determine which attributes were critical in the classification process. We validated the results using 5-fold cross validation, as well as through three neural network classifiers – the feed-forward error back-propagation network, the Radial Basis Function (RBF) network, and the Local Transfer Function Classifier (LTF-C). This validation was performed in a two-stage process (for Multi Layered Perceptron (MLP) only). We first trained and tested the neural networks on the full dataset (containing 20 attributes), we then performed attribute dimensionality reduction using rough sets, and then tested the neural networks with the remaining attributes generated from the rough sets analysis. We used the LTF-C neural network and a modified k-NN classifier as independent measures of the accuracy of the results we obtained with rough sets. To our knowledge, this is the first paper reporting this type of analysis within the context of EEG analysis.

2 Data Acquisition and Pre-processing

We have used publicly available datasets described in Andrzejak et al. [1]. The complete data set consists of five sets (denoted A–E) each containing 100 single-channel (100 electrodes) EEG recordings of 5 separate patient classes. For this study we focused on sets labelled A and E in [1]: the normal and epileptic

seizure session recordings. These segments were selected and cut out from continuous multi-channel EEG recordings (i.e. 23.6 seconds of recording time) after visual inspection for artifacts, e.g., due to muscle activity or eye movements. In Figure 2, the panel labelled 'Set A' corresponded to the normal class and panel 'Set E' is an example of epileptic seizure class. All EEG signals were recorded with the same 128-channel amplifier system, using an average common reference. The data were digitised at 173.61 Hz using 12-bit resolution. The data was band-pass filtered at 0.53–40 Hz (12dB/oct). Each EEG dataset consisted of 4,096 data points and a rectangular window of 256 discrete data points (16 windows per electrode) was selected.

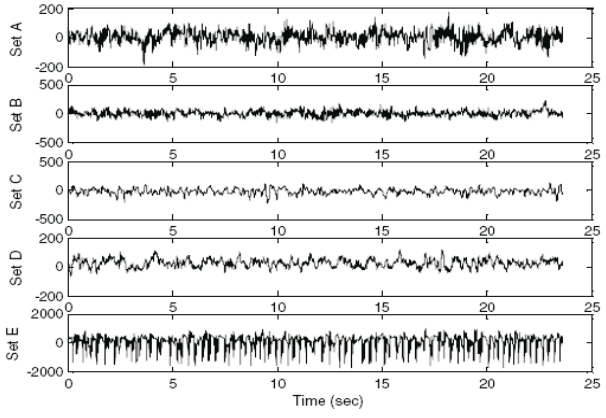

Fig. 1. Examples of five different sets of EEG signals taken from different subjects

Selection of suitable wavelet and the number of decomposition levels is very important in analysis of signals using the DWT. The number of decomposition levels is chosen based on the dominant frequency components of the signal. The levels are chosen such that those parts of the signal that correlate well with the frequencies necessary for classification of the signal are retained in the wavelet coefficients. In the present study, since the EEG signals do not have any useful frequency components above 30 Hz, the number of decomposition levels was chosen to be 4. Thus, the EEG signals were decomposed into details D1–D4 and one final approximation, A4. Usually, tests are performed with different types of wavelets and the one which gives maximum efficiency is selected for the particular application. The smoothing feature of the Daubechies wavelet of order 2 (db2) made it more appropriate to detect changes of EEG signals. Hence, the wavelet coefficients were computed using the db2 in the present study. Figure 2 shows approximation (A1–A4) and details (D1–D4) of an epileptic EEG signal.

The detail wavelet coefficients (D1–D4) at the first, second, third, and fourth levels (129+66+34+18 coefficients) and the approximation wavelet coefficients (A4) at the fourth level (18 coefficients) were generated. We have 100 electrodes and 16 windows per electrode, yielding a total of 1,600 segments per class.

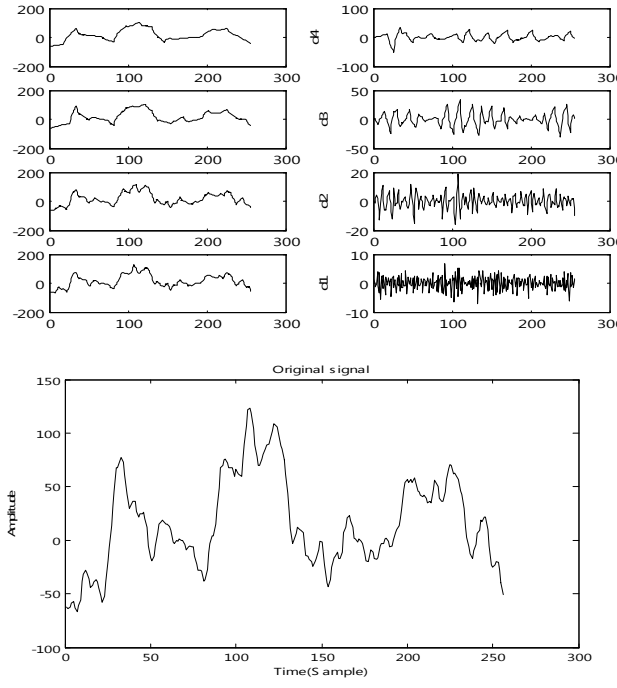

Fig. 2. Approximate and detailed coefficients of EEG signals taken from epileptic subject

The extracted wavelet coefficients provide a compact representation that shows the energy distribution of the EEG signal in time and frequency. For each set of detail coefficients (D1-D4) and the approximation wavelet coefficients (A4) we calculate four values: maximum, minimum, mean and standard deviation. That yields a final set of 20 real valued attributes for each class (following the ideas from [6]).

3 Classification Methods

In this study, we employed a combination of neural network based classification algorithms in conjunction with rough sets. Initially, we trained an MLP and RBF neural network to classify the 2 EEG classes (A and E). This was accomplished through the standard back-propagation learning algorithm and the orthogonal least squares algorithm was used to train the RBF network. The LTF-C algorithm uses a modified version of the RBF algorithm, which is explained in detail below. We next briefly describe the rough sets and neural network based classifier, before presenting the main results of this study.

3.1 Rough Sets

Rough Set (RS) theory is a main topic of this conference and that our study uses only some very basic and well known RS tools we restrict ourselves to introducing

only these RS notions that we further make use of. As the capacity of this paper is very limited, we make only a basic description of these concepts.

The first step in the process of mining any dataset using rough sets is to transform the data into a decision table. In a decision table (DT), each row consists of an observation (also called an object) and each column is an attribute, one of which is the decision attribute for the observation. In our case the decision table consists of 8000 rows and each row contains a vector of 20 numbers (values of 20 numerical conditional attributes) labelled with one of 5 decision values (A-E). Objects that share the same decision value are said to belong to one *decision class*. Attributes other than decision will be referred to as *conditional attributes* or simply *conditions*.

Luckily enough, thanks to the fact that data table is generated by a controlled pre-processing algorithm, we have no missing values and no errors which is not commonplace in medical data sets. The table is consistent, i.e. there are no two rows that have the same conditional part and different decisions.

One of most characteristic features of our data set is that all conditional attributes are numeric (floating point numbers). In order to apply some of the RS methods to such data table one have to perform *discretisation*.

Discretisation refers to partitioning attributes into intervals – tantamount to searching for "cuts" in the range of attribute. All values that lie within a given range (between two cuts) are mapped onto the same value, transforming interval into categorical data. In this study we apply discretisation method based on Maximal Discernibility (MD) heuristics that makes use of the core RS notion of *discernibility* between decision classes. Details of this method and its implementation in RSES are given in [2,3].

The ultimate goal we want to achieve with our RS toolkit is to construct a *classifier* – that is a procedure which when given an unlabelled object is capable of assigning a proper decision value. In particular, we will be dealing with classifiers that are based on *decision rules*, i.e., formulae of the form:

$$(a_i = v_i) \wedge \ldots \wedge (a_i = v_i) \Rightarrow (d = v)$$

where atomic sub-formula $(a_i = v_i)$ is called *descriptor* or *condition*. We say that rule r is *applicable* to an object, or alternatively, the object *matches* rule, if its attribute values satisfy the premise of this rule. With the rule we can associate some numerical characteristics derived from the underlying data table DT. $Supp(r)$ is equal to the number of objects from table for which rule r applies correctly, i.e., the premise of rule is satisfied and the decision given by rule is similar to the one preset in decision table. $Match(r)$ is the number of objects in the table for which rule r applies in general. Analogously the notion of matching set for a rule or collection of rules may be introduced (see [3]).

The notions of matching and supporting set are common to all classifiers, not only decision rules. For a classifier Cl we will denote by $Supp(Cl)$ the set of objects that support classifier, i.e., the set of objects for which classifier gives the answer (decision) identical to that we already have. Similarly, $Match(Cl)$ is the set of objects that are recognized by Cl. Support and matching make it possible

to introduce two measures that are used in our study for classifier scoring. These are *Accuracy* and *Coverage*, defined as follows:

$$Accuracy_{DT}(Cl) = \frac{|Supp(Cl)|}{|DT|}; \quad Coverage_{DT}(Cl) = \frac{|Match(Cl)|}{|DT|}$$

where $|DT|$ denotes number of objects in our data table.

3.2 MLP, RBF, and LTF-C Networks

The Multilayer Perceptron Network (MLP), which has the ability to learn and generalise, smaller training set requirements, fast operation, ease of implementation and therefore most commonly used neural network architectures, have been adapted for describing the alertness level of arbitrary subject. We have used in this case, the classic gradient descent learning scheme for the training of this particular network.

The second classification scheme utilised here is a Radial Basis Function Network (RBF) scheme. RBF networks train rapidly, usually orders of magnitude faster than MLP, while exhibiting none of its training pathologies such as paralysis or local minima problems. Such a system consists of three layers (input, hidden, output). The activation of a hidden neuron is determined in two steps: The first is computing the distance (usually by using the Euclidean norm) between the input vector and a centre c_i that represents the i^{th} hidden neuron. Second, a function h that is usually bell-shaped is applied, using the obtained distance to get the final activation of the hidden neuron. In this case the Gaussian function $G(x)$ was used. The parameter σ is called unit width and is determined using the heuristic rule "global first nearest-neighbour". The activation of a neuron in the output layer is determined by a linear combination of the fixed nonlinear basis functions, i.e. here $\phi_i(x) = G(\|x - c_i\|)$ and w_i are the adjustable weights that link the output nodes with the appropriate hidden neurons. The orthogonal least squares (OLS) method has been employed as a forward selection procedure that constructs RBF networks in a rational way. The algorithm chooses appropriate RBF centres one by one from training data points until a satisfactory network is obtained.

The last classification scheme utilised in this study was based on the Linear Transfer Function Classifier (LTF-C) scheme in the version implemented in RSES. LTF-C (cf. [12]) is a neural network solving classification problems. Its architecture is very similar to this of RBF – the network has a hidden layer with Gaussian neurons connected to an output layer of linear units. The number of inputs corresponds to the number of attributes while the number of linear neurons in output layers equals the number of decision classes. There are some additional restrictions on values of output weights that enable to use an entirely different training algorithm and to obtain very high accuracy in real-world problems. The training algorithm of LTF-C comprises four types of modifications of the network, performed after every presentation of a training object. Namely the network can: change positions (means) of Gaussians in hidden layer, change widths (deviations) of Gaussians separately for each hidden neuron and attribute, insert new hidden neurons, and remove unnecessary hidden neurons.

As one can see, the LTF-C structure is dynamical. The training process starts with an empty hidden layer, adding new hidden neurons when the accuracy is insufficient and removing the units which do not positively contribute to the calculation of correct network decisions. This feature of LTF-C enables automatic choice of the best network size, which is much easier than setting the number of hidden neurons manually.

4 Results

As previously indicated, we initially trained the MLP, RBF, and LTF-C networks to classify the two EEG time series using 20 inputs and a binary decision class (either A or E) for the output. We used a 50/50 rule where 50% of the data (half of 3,200 objects in total) were used for training and the rest was used for testing. This train-and-test scheme was selected for the sake of obtaining results that are directly comparable with those existing before. More sophisticated testing schemes are planned for the follow-up of the currently presented study. The results in Table 1 represent the training and testing accuracy of the neural network classifiers.

Table 1. The classification accuracy of the MLP, RBF and LTF-C on the 2 classes EEG testing data. The values in parentheses represent training accuracy.

	MLP	RBF	LTF-C
Class A	(95.1) 94.3%	(96.1)95.2%	(100) 99.8%
Class E	(96.5) 93.8%	(97.3)96.8%	(100) 99.8%

Relatively high quality obtained by neural network models without any special fine-tuning was an indicator that they may exist a simple regularity in the data that makes it possible to discern between classes A and E. To find out what kind of regularity may that be and what possible use we may have of it for our study, we turned to rough set methodology. Using some typical rough set based algorithm implemented in the Rough set Exploration System (RSES, see [11]) we started to search for a set of simple description rules discriminating between classes A and E. Since the considered data set has all conditional attributes represented numerically, the discretisation procedure has been applied. The use of discretisation, reduct and rule calculation has resulted in a very interesting and somehow surprising results. Since rough set methods we have used concentrate on maximising discernibility and minimising (reducing) the dimensionality of derived model, they tend to provide a concise description. In our case this these descriptions have in fact become ultra-compact. As it turns out the DWT data contains several attributes that have very high discriminative abilities with regard to classes A and E. Some attributes make it possible to construct a set of rules with only a single condition (on this attribute) and almost perfect accuracy on the entire data set. The table below presents a simple summary of the identified attributes.

Table 2. Qualities of selected single-attribute rule sets for classes A and E

Attribute	No. of rules	Accuracy	Attribute	No. of rules	Accuracy
Max. D2	2	99.8%	Min. D4	2	99.9%
Min. D2	10	99.6%	Max D4	2	100%
St. dev. D2	14	99.2%	Min. A4	6	99.9%
Min D3	8	99.9%	Max A4	10	99.7%
Max D3	2	99.9%	St. dev. A4	2	99.9%
St. dev. D3	2	99.8%			

Note that a single attribute Max D4 is sufficient to create a perfect classifier if we only discern between classes A and E. Such classifier would consist of only two 100% correct decision rules:

```
If Max D4 < 173.1838 then Decision = A;
If Max D4 > 173.1838 then Decision = E;
```

Also, the results in table show that basically for all listed attribute there is only a handful of outliers that prevents each of them from being 100% correct. These errors are in fact almost entirely dependant on the random split. Inasmuch as experiments have shown that classes A and E are easily discernible we decided to investigate how different is the situation if we not only focus on these two classes. In the more realistic setting we will not know in advance whether the given measurement comes from the patient from class A/E or any of the other (B,C,D). Therefore, even if we focus on A and E, we still have to be able to discern between A, E and the rest. If it proves possible to do so, then we may attempt to construct a decision support system that at the same time tells us whether a given data example is in our focus group (A+E) or not and whether it is a healthy (A) or epileptic (E) patient.

We have staged an experiment with use of various classifiers implemented in RSES in order to verify how well can we differentiate classes A and E from the rest. The training and verification in this experiment was performed on the entire 8000 rows (all classes). As in the previous experiments the data was randomly halved in order to create training and test samples. Once again we have used LTF-C and a rule-based RS classifier from RSES (see see [11] for details). For reference we have also applied a modified *k-Nearest Neighbours* (k-NN) classifier from RSES. The k-NN classifier is known to provide very good results, however, without giving an explanation. The results were rather encouraging. Classes A and E are not only well discernable from each other, but also significantly different from the rest of data set (B+C+D). Table 3 presents overview of the results. Since the methods used are always 100% accurate on training sample we only show average results on test cases.

We have also performed an initial attempt to construct classifier for all 5 types of patients. As expected, the performance of such classifier is significantly lower (on average more than 10%). Rule based methods tend to have problems with unseen cases (low coverage). It is quite visible that the three sets B, C, and D

Table 3. Classification results for classes A, B+C+D, and E on test set

Classifier	Classifier details	Avg. accuracy	Avg. coverage
Modified k-NN	Reference result	99.4%	100%
LTF-C	Neural network	98%	100%
Decision rules (all)	Up to 13000 rules	97.2%	99%
Decision rules (LEM)	200-220 rules	98.7%	95.2%
Decision rules (LEM)	270-290 rules	98.5%	96.5%

are more complicated and vague than A and E. Constructing a proper classifier for the entire dataset is on our task list for next stage of this study.

Notice: *All the experimental results presented in this section have been averaged over several repetitions in order to avoid presentation of random, insignificant and non-replicable outputs.*

5 Conclusions

In this study, we examined the difference(s) between normal and epileptic EEG time series. After extracting the DWT coefficients, we used four measures of dispersion to as attributes for subsequent classification. The results of the neural network (MLP and RBF) classification were high, comparable to other published results on similar datasets. We then sought to perform dimensionality reduction through the rough sets paradigm. The results from this analysis indicated that the statistical attributes (20 in all) contained a considerable amount of redundancy. Rough sets was able to reduce the dimensionality of the attributes to a single one – Max D4 (see Table 2). With this single attribute, only 2 rules were generated which provided 100% classification accuracy. This result was confirmed with independent methods such as a modified k-NN and the LTF-C. We also used this single attribute to re-test the original MLP network that were trained on the full dataset and the resulting classification accuracy was not reduced using the single attribute (Max D4). The resulting classification accuracy was in fact somewhat higher (98%) than when trained on the full set of attributes.

These results indicate that there is the potential for considerable redundancy with attribute selection in this particular domain. Our results indicate that preprocessing the data using rough sets is an effective way of eliminating this potential redundancy in attribute selection. With a minimal attribute set, one can then begin to find more exacting correlations between the behaviour of attributes and the underlying phenomenon. In this particular case, Max D4 was the primary attribute for distinguishing normal from epileptic EEG recordings. The 4^{th} level represents the most refined sampling performed on this data. We are planning to investigate this result in future work – but in this instance one can hypothesise that the increased sampling rate provides significant information. The 'Max' represents the maximal value within the window – and one could reasonably explain this result in terms of the reported spiking that occurs in epileptic seizures.

The results from this preliminary study will be expanded to include a more complete range of pathologies. In this work, we focused on the extremes that are found within the EEG spectrum – normal and epileptic time series. These two series were chosen as they would more than likely lead to the maximal dispersion between the 2 signals and be amenable for training of the classifiers. In the next stage of this research, we have datasets that are intermediate in the signal changes they present. This will provide a more challenging set of data to work with – and will allow us to refine our learning algorithms and/or approaches to the problem of EEG analysis.

Acknowledgements. The authors wish to thank the authors of EEG dataset which is publicly available at
http://www.meb.uni-bonn.de/epileptologie/science/physik/eegdata.html.

Marcin Szczuka is partly supported by Polish grant 3T11C02226.

References

1. Andrzejak R. G., Lehnertz K., Mormann F., Rieke C., David P., Elger C. E., Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity. *Phys. Rev. E*, 64, pp. 1-8, 2001.
2. Bazan J., Nguyen H.S., Nguyen S.H., Synak P., Wróblewski J., Rough set algorithms in classification problems. *Studies in Fuzziness and Soft Computing* vol 56,. Physica-Verlag, Heidelberg, pp. 49-88, 2000.
3. Bazan J., Szczuka M., The Rough Set Exploration System. Transactions on Rough Sets III, **LNCS** 3400, Springer, pp.37-56, 2005. http://logic.mimuw.edu.pl/∼rses
4. Folkers A., Mosch F., Malina T., Hofmann U. G., Realtime bioelectrical data acquisition and processing from 128 channels utilizing the wavelet-transformation. *Neurocomputing*, 52–54, pp. 247-254, 2003.
5. Glover Jr. J.R., Raghaven N., Ktonas P.Y., Frost Jr. J.D., Context-based automated detection of eleptogenic sharp transients in the EEG, *IEEE Tans. Biomed. Eng.*, 36(5), pp. 519-527, 1989.
6. Guler I., Ubeyli E.D., Application of adaptive neuro-fuzzy inference system for detection of electrocardiographic changes in patients with partial epilepsy using feature extraction. *Expert Syst. Appl.*, 27(3), pp. 323-30, 2004.
7. Nigam V.P., Graupe D.A, Neural-network-based detection of epilepsy, *Neurological Research*, vol. 26:1, pp. 55-60, 2004.
8. Ningler M., Stockmanns G., Schneider G., Dressler O., Kochs E.F., Rough Set-based Classification of EEG-signals to detect Intraoperative Awareness, Proc. of RSCTC2004, **LNAI** 3066, Springer, pp. 825-834, 2004.
9. Petrosian A., Prokhorov D., Homan R., Dashei R., Wunsc, D., Recurrent neural network based prediction of epileptic seizures in intra and extracranial EEG. Neurocomputing, 30, pp. 201–218, 2000.
10. Pradhan N., Sadasivan P. K., Arunodaya G. R., Detection of seizure activity in EEG by an artificial neural network. *Computers and Biomedical Research*, 29, pp. 303–313, 1996.
11. Weng W., Khorasani K., An adaptive structure neural network with application to EEG automatic seizure detection. Neural Networks, 9, pp. 1223-1240, 1996
12. Wojnarski M., LTF-C: Architecture, training algorithm and applications of new neural classifier. *Fundamenta Informaticae*, 54(1), pp. 89-105, 2003.