

# Neural Networks and Rough Sets: A comparative study on data classification

**Renato J. Sassi**  
Department of Electronic  
Systems Engineering  
University of Sao Paulo, Brazil  
(e-mail: sassi@lsi.usp.br)

**Leandro Augusto da Silva**  
Department of Electronic Systems  
Engineering  
University of Sao Paulo, Brazil  
(e-mail: leandro.augusto@ieee.org)

**Emilio Del Moral Hernandez**  
Department of Electronic Systems  
Engineering  
University of Sao Paulo, Brazil  
(e-mail: emilio\_del\_moral@ieee.org)

**Abstract** - *This paper addresses a contrastive study between Neural Networks and Rough Sets on data classification. The experiments were carried out using the Iris database, of public domain, to evaluate the classification. The confusion matrix method was used to evaluate the performance of these classifiers. With these contrastive experiments, we investigated the capacity of each classifier for application in a potential application on knowledge extraction in databases. In this experiment the results indicate that the Neural Networks classifier, except SLP, presents significant superiority on Rough Sets classifiers.*

**Keywords:** Rough Sets, Neural Networks, Data Classification, Reduction of Attribute.

## 1 Introduction

The main basic function executed by human brain is the classification. The human can analyze objects using some characteristics to perceive the differences and similarities.

So, is possible classify animals as friendly or dangerous, healthful plant for be eat or not, etc. In all moments of your life, the human impose classification among two or more object.

Several techniques are used to data classification like statistics methods, Neural Networks and Rough Sets. In this work the contrasted study the use Neural Network or Rough Sets in data classification. Due this capacity, two techniques are used in knowledge extraction in database.

A Neural Network trained with a specific set of examples can be used to classify elements that define if some data characteristic belong to a class or no. Thus, the Neural Network has been used in commercial applications to classify customer, fraud detection, handwriting recognition, datamining, medical diagnostic, speech recognition, and othes [1].

The Rough Sets study was proposed by Pawlak [2] in the early 1980's with a mathematic approach to uncertainty analyze in data. The proposed study an investigation of that can be indiscernible (similar) due the limited information available about them. The use of Rough Sets in knowledge extraction field can be proved by growing applications

number and publication [3,4,5,6,7,8]. According Pawlak [9], one of the main advantages of Rough Sets is the non require information a priori about the data. The Rough Sets characteristics are [10]:

- Characterization of set of objects in terms of attributes values.
- Finding dependencies (total or partial) between attributes.
- Reduction of superfluous attributes (data).
- Finding the most significance attributes.
- Decision rule generation to data classification.

In this work a comparative study on data classification using Rough Sets and different Neural Networks architectures (SLP, MLP, RBF and SOM) is proposed.

The content of this paper is organized as follows. In Section 2, na early introduction about the Neural Networks architectures used in contrastive study. In Section 3 a detailed explanation of Rough Sets theory in classification context. Experimental methodology and results obtained are provide in Section 4 and 5. Finally, conclusion is drawn in Section 6.

## 2 Neural Network

The Neural Network are ample effectives to pattern learning from non-linear data, imperfect, with noises and also composite with mistakes examples. Applications examples using Neural Network are: pattern recognition (images, text, speech); forecast series temporal; and others [11]. In general, the Neural Network is composed with neurons in layers (one or more, depending of architecture) connected. In the most of architecture these connections are associated to a weight which storage the data knowledge and are used to ponder the input data in each neuron from network [11],[12]. This characteristic become the Neural Networks a interesting tool in knowledge extraction [13].

The main Neural Networks architecture is SLP (Single Layer Perceptron), MLP (Multi Layer Perceptron), RBF (Radial Base Function) and SOM (Self-Organizing Maps). The fundamental characteristics structural and functional will be early discussed next.

## 2.1 Single Layer Perceptron (SLP)

The SLP is assembling by neurons in two layers, input and output, fully coupled. The involve learning in its training is iterative and supervise. For iteration the training patterns and its respective classes are presented to SLP. To receive one training pattern this forward to output layer. In this moment the error between the class desire and the output value given to SLP. So, if this error is minor the maximum error accepts (previously defined) the learning is stopped else the learning go on and the error measure is used to adjust the synaptic weight, which initially can be initialized random, and so another training pattern is present. Thus, the algorithm continues to minimum error reaching or also a maximum iteration number, defined in initial of process too The algorithm in question is call least mean square (LMS) and a complete explanation is shown in Haykin [11]. In the end learning, the synaptic weights are freezes and can be used to a classification, pattern recognition, etc. However the SLP not well-work with when the patterns are not linear separable [11].

## 2.2 Multi-layer Perceptron (MLP)

The MLP as well as SLP, has architecture in layers. However the MLP can have a hidden layer (one or more) situated between input layer and output layer. The neurons of layers are fully connected. The involve learning in its training is iterative and supervise, as well as SLP. The difference from SLP, beyond hidden layer, is that the measured error after iteration is used to adjust the synaptic weight (see Haykin to a detail explanation [11]). This process progress to minimum error reaching or a maximum iteration number, both initial defined. The algorithm quickly previous discuss is call error back-propagation [11].

The MLP is the most Neural Network and it is used in several applications like classification, pattern recognition, function estimation and others. Differently SLP, the MLP well-work when the classes non-separable linear [11].

## 2.3 Radial Base Function (RBF)

As well as SLP and MLP, the RBF (*Radial Base Function*) is architecture in layer. However the RBF layers are three with different function. The input layer receive the input patterns set which are forward propagation, in the second later, unique hidden, is responsible by non-linear application of input patterns, in general of high dimension, because radial function and the output layer which is a linear transformation. The argument for a non-linear transformation to a linear transformation is that the pattern order in a high dimension have more probability to be separable linear and so the classification is more correct. The RBF learning is unsupervised, in other words, is not necessary shown the patterns classes during the training.

The algorithm considerate in learning is better discussed in Duda et. al [14].

As well as MLP, the RBF is used to classify patterns with linear separable and non-linear separable [11],[14].

## 2.4 Self-Organizing Maps (SOM)

The SOM or Kohonem Map, in one more neural network in layer, input layer and output layer that are fully connected. The SOM learning is unsupervised. As the RBF, the SOM not depend of priory pattern classes' information. During the training the patterns self-organizing in a map (or grid) and the position information of grid can be used to labeling the data what become the use of SOM in classification application. The SOM has as characteristic capture similarities and correlations from patterns. This is realized through a competition between neurons during the training which the winning must be active as a of a specific pattern from training. Thus, from pattern training set, the algorithm pick a pattern from set of n-dimension and provide a output (map) that better represents this set [11], [14] e [15].

The SOM beyond used in classification task, he is most used in data mining [15].

# 3 Rough Sets

The Rough Sets studies propose identify of that patterns can be indiscernible (similar) despite have an information limited about them. Rough Sets has properties that provide reduce irrelevant attributes (reduce pattern dimension) through of knowledge of all training patterns without the classes, which are call in Rough Sets theory as Information System (IS). This properties is call as reduct which is realizable that the pattern information (like what class belong the pattern) be sustained, in others words, without change the pattern knowledge representation [2]. In this context, the patterns that can not specified from set available are characterized by two concepts: lower approximation and upper approximation, which are discuss in this section yet.

## 3.1 Approximation space

An approximation space is an ordered pair  $A = (U, R)$ , where:

- U is a nonempty set, designated universe set;
- R is an equivalence relation about U, designated as indiscernibly relation. Given  $x, y \in U$ , if  $xRy$  then x and y are indiscernible in A, in other words, the equivalency class defined by x is the same that defined by y, i.e.,  $[x]R = [y]R$ .

In a IS, each object (pattern) has a set of attributes (see TABLE I). These attributes have the same characteristics but the values are different.

In this way, a IS is an ordered pair  $S = (U, A)$ . The elements where U is a finite set and non-empty of elements call attributes. The elements from universe will be call as

objects. Each attribute  $a \in A$  is a total function  $a : U \rightarrow Va$  where  $Va$  is the values set possible to  $a$  attribute (range of values).

In many cases is important the classification of objects considering a decision attribute (D) that notify the decision to take. Can assume that a IS with the D is a Decision System (DS).

DS is any SI from  $A = (U, A \cup \{d\})$ , where  $d \notin A$  is the D. The A elements are call of conditional elements or only conditional.

Sepal Length	Sepa Width	Petal Length	Peta Width
5.1	3.5	1.4	0.2
7	3.2	4.7	1.4
5.8	2.7	5.1	1.9

TABLE I. The columns are the attributes and the lines, exception first are the objects. For this IS sample (from iris database) we have three objects where each have four attributes.

Sepal Length	Sepa Width	Petal Length	Peta Width	Classes
5.1	3.5	1.4	0.2	setosa
7	3.2	4.7	1.4	versicolor
5.8	2.7	5.1	1.9	virginica

TABLE II. The Decision System is equal the IS, the difference is the last column with the classes value (attribute D).

### 3.2 Indiscernibility Relation

The main concept involving in RS is the indiscernibility relation which in general is associate with a attributes set.

If the relation exist between two objects this mean that all values of its attributes are equal in relation to considerate attributes. However these can not be discernible (distinguishable) between itself considering these attributes.

First time, the Rough Sets define the indiscernibility relation, where is defined the redundant attributes or dispensable.

A Decision System (DS) (see TABLE.II) is the IS with the pattern classes, or in Rough Sets context, the decision attribute. In a DS all knowledge of patterns set is expressed and it can be redundant at least of two forms [16]:

1. The indiscernible objects can be represented several times.
2. Some attributes can be superfluous.

A binary relation  $R \subseteq X \times X$  which is reflexive (i.e., an object is relational with itself  $xRx$ ), symmetric (if  $xRy$  then  $yRx$ ) and transitive (if  $xRy$  and  $yRx$  then  $xRz$ ) is call equivalence relation.

The equivalence class of one element  $x \in X$  consist of all objects  $y \in X$  for which  $xRy$ .

For each subset of attributes  $B \subseteq A$  in SI  $A = (U, A)$ , a relation of equivalence  $IND A (B)$  is associate, call relation indiscernibility relation defined as:

$$INDA(B)=\{(x,y) \in U^2 \mid \forall a \in B, a(x)=a(y)\}$$

Where  $INDA(B)$  is call indiscernibility relation. The set of all equivalence classes in relation  $IND A (B)$  is denote by  $U/IND A (B)$ .

With the indiscernibility relation concept is easy define redundant attributes. If attributes set and its subset define the same indiscernibility relation, in other words, the elements that are indiscernible (elementary concepts or basic granules) of both relations are identical so, any attribute that belong to subset and not belong to attribute set is denoted as redundant.

An attribute set that not have redundant attribute is call minimum set or independent. A P attribute set is a reduct of another attributes set Q, if P is minimum and the relations of indiscernibility relation, defined by P and Q are the same, in other words, the elementary concepts defined by P and Q are identical [2].

### 3.3 Approximations

In this way, Rough Sets provide a simple form to treat with the uncertainty. For each X, the bigger definable set (finite union of elementary concepts) restrained in X and the lesser definable set that contain X are computed. The first set is called X-lower approximation and the second X-upper approximation [2]. In this work the approximations will be designed respectively as:  $B\_X$  and upper approximation as  $B^-X$ .

The  $B\_X$  and  $B^-X$  of an object set  $X \subseteq U$  in despite an attributes set  $B \subseteq A$  (define an equivalency relation in U) can be defined in terms of the equivalency relations classes as:

$$B\_X = \cup \{E \in U/IND(B) \mid E \subseteq X\}$$

$$B^-X = \cup \{E \in U/IND(B) \mid E \cap X \neq \emptyset\}$$

The  $B\_X$  is an elements set that can be classified with assurance as a member of X set using the attributes set B. In a similar manner, the  $B^-X$  are the elements that can be classified as set X member using the attributes set B.

The boundary region have the objects that can not be classified with assurance as belong or no to X using the attributes set B. The boundary region are composed by U elements that belong  $B^-X$ , but no belong to  $B\_X$ .

The  $BNB (X) = B^-X - B\_X$  is call B-border X. A set is call as rough (imprecise) if the border region is non-enough. A set is call crisp (precise) if border region is enough.

### 3.4 Reduction of Attributes

There is possibility of not all the attributes be used to discern classes with the purpose to recognition who are the attributes important to determine the classification emerge the reduct.

A reduct is a minimum attributes subset conditional that possibility takes the same decisions of complete set. In other words, a reduct is minimum attributes subset (reduct) that retain the decision attributes dependence degree to conditional attributes [2,10,16].

An  $a$  attribute is dispensable or superfluous or redundant in  $B \subseteq A$  if  $IND(B) = IND(B) - \{a\}$ , contrary case is indispensable in  $B$ . If all attributes  $a \in B$  are indispensable in  $B$ , so  $B$  is call orthogonal. A  $B$  reduct is a attributes set  $B' \subseteq B$  such that all attributes  $a \in B - B'$  are dispensable an  $IND(B') = IND(B)$ .

Next is shown the Rough Sets Reduct from Iris database

Reduct	Support	Length
{Sepal Length, Petal Length, Petal Width }	100	3

The support 100 indicate that this reduct generated precise rules (strong) and so can be take a correct decision using only this reduct because define unique decision to conditional attributes.

### 3.5 Rules

The rules represent dependents in patterns set and knowledge extracted which can be used to classify new patterns. The rules are generation thought union of attributes conditional values from patterns used in reduct. With the rules can be do inferences to a precision classification. A rule example to TABLE II is:

**SE SepalLength AND PetalLength AND PetalWidth =>  
Class(Iris-setosa)**

## 4 Computational Experiments

For experimental results we use a public database and to measure performance of classifiers we use the confusion matrix. Next a briefly discussion about these.

### 4.1 Database

For computation experiments will be used a public database [17]. The database call iris, is formed by patterns from three different classes (Setosa, Versicolor, Virginica) referents the plant Iris. For each class a total of 50 samples. Each sample is composed by four numeric continuous

attributes as measure of weight and length of sepal and measure of weight and length of petal. Accord with database information [17] the Setosa class is linear separable from others, so, Versicolor and Virginica are not linear separable between itself.

### 4.2 Measure Performance

For performance measure of classifiers will be used the confusion matrix. This matrix is so used to analyze classifiers performance.

The confusion matrix is a table where is possible if the tests patterns are being correctly or incorrectly classified. The structure of matrix confusion is shown in TABLE III.

The first lines represent the  $n$  pattern classes to be classified and the column represents the classifier response (patterns classified). Thus, the principal diagonal of confusion matrix reveals the classifier hit. As experimental results, it will be shown by sum of principal diagonal.

## 5 Obtained Results

For the experiments we separated two set from database, one to training and another to test. For this a random chooses from 150 samples. Thus, after the random choose we separated 100 samples as training patterns and 50 samples as test patterns. A total of 20 experiments with different random choose of training pattern and test patterns was conducted. For a each choose, experiments with all classifiers contrasted in this work were conducted and in the final, the mean and the standard deviation of sum of principal diagonal of confusion matrix for each classifier is calculate as shown in TABLE III. Another results contrasted, second column TABLE III is the mean training time of classifiers in 20 different experiments. All experiments were realized in a Pentium 4 with 256M of RAM memory and 2.0GHz. In relation classifiers configuration, this will be discuss next.

		Classified Patterns		
		$C_1$	$C_2$	$\dots C_n$
Patterns to be classified	$C_1$	X Classified as $C_1$ (correct)	Y Classified as $C_2$ (incorrect)	Z Classified as $C_n$ (incorrect)
	$C_2$	X Classified as $C_1$ (incorrect)	Y Classified as $C_2$ (correct)	Z Classified as $C_n$ (incorrect)
	$\dots C_n$	X Classified as $C_1$ (incorrect)	Y Classified as $C_2$ (incorrect)	Z Classified as $C_n$ (correct)

TABLE III. Confusion matrix. Where  $C_1, C_2 \in C_n$ , represents the patterns classes.

As discussed in 2.1 the SLP classify two classes with linear separated, but in this experiments the Iris samples are separating in three classes we conduce the experiments as following. The training and the test were realized contrasting the classes two to two. Thus, contrasting the Setosa sample with Versicolor samples, Setosa samples with Virginica samples and Versicolor samples with Virginica samples. The classification result is to class that shown more correct classification of a specific class (*arg\_max*). The architecture of SLP has 4 neuron in input layer and three in output layer and the stopping condition was 2.000 maximum numbers of iterations, minimum error 0.01. The MLP architecture has 4 neurons in input layer, one hidden layer with 5 neurons and 3 in input layer. To choose the number hidden layer as well as the neuron number on hidden layer, experimental test as proposed in Haykin [11] were conducted. The training parameters to MLP, 2.000 maximum iteration number minimum and error 0.01. The RBF architecture with 4 input layer and 3 neurons in hidden layer and 3 neurons in output layer. As transference function in hidden layer we used the Gaussian function with base width equal to *standard deviation = dmax / sqrt(2\*m<sub>i</sub>)* where *m<sub>i</sub>* is the center number estimated using the K-means (more details see Duda et al. [14]), where K represents the center number, 3, and *dmax* is the maximum distance between the e choose centers.

	mean and the standard deviation of sum of principal diagonal of confusion matrix to 20 experiments	Training time mean (in seconds)
<i>SLP</i>	65.67 ± 2.52	101
<i>MLP</i>	96.16 ± 1.59	150
<i>RBF</i>	96.71 ± 1.88	0.01
<i>SOM</i>	98.33 ± 0.00	0.27
<i>Rough Set</i>	76.78 ± 0.05	*

TABLE IV. Table of Results

\* the Rough Sets classifier has not execution time, but operation time because to each step, a manual operation is necessary.

In these three experiments SLP, MLP and RBF, the functions were developed using MATLAB software.

For the SOM, explained next, the toolbox [18] was used. The SOM architectures 4 neuron in input layer and an output grid of 8x2, 16 neurons in output layer. On the training were spend 1000 iteration with learning rate equal 0.1 (order phase) and 500 iterations with learning rate equal 0.01 (convergence phase).

For Rough Sets experiments the Rosetta toolbox [19] was used. Initially the sample continuous values must be discrete, for this the used method was the Boolean Reasoning Algorithm – RSES. After this, the reduct are

generated using genetic algorithms. Consequently the rules are generated. For the classification we used the Standard-Vote Algorithm. All the function previously cited (Boolean Reasoning Algorithm – RSES, Genetic Algorithms, Standard-Vote Algorithm) are part from Rosetta toolbox.

## 6 Conclusions

The experimental results by neural classifiers are relation straight with linear separable of classes. Thus, the classifier neural SLP result is was inside of the waited one, in other words its result is the inferior. On another hand, the non-linear neural classifiers (MLP, RBF and SOM) present superior result to SLP, but equivalent.

In relation the Rough Sets results it is better the SLP, but worst than others neural classifiers. The hypotheses for this results is that the iris database samples are continuous values and need be discrete which difficult the attributes characterization.

As future work, we intend combine one Neural Network architecture with Rough Sets, hybrid architecture. With this, we intent explore the main characteristic of each classifier for a knowledge extraction database application.

There are works that propose the hybrids architectures [20,21].

## 7 References

- [1] J. P. Bigus, “Data Mining with Neural Network: Solving Business Problems from Applications Development to Decision Support”, McGraw-Hill.1996.
- [2] Z. Pawlak, “Rough Sets”, International Jornal of Computer and information Sciences, 1982, pp.341-356.
- [3] T. Beauboef and F.E. Petry, “A Rough set foundation for spatial data mining involving vague regions. Fuzzy Systems”, 2002. FUZZ-IEEE’02. Proceedings of the 2002 IEEE International Conference on, volume 1, 12-17 May 2002.
- [4] S. Tsumoto, “Knowledge Discovery in Medical Databases based on Rough Sets and attribute-oriented generalization” IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on Fuzzy Systems Proceedings, 1998, volume 2, 4-9 May 1998, pages: 1296-1301.
- [5] H. Xiaohua and N. Cercone, “Mining knowledge rules from Databases: a Rough set approach” Proceedings of the Twelfth International Conference on Data Engineering, 1996. 26 Feb.-1 March 1996, pages: 96–105.
- [6] W. Xiao-Ye and O.W. Zheng, “Stock Market Time series Data Mining based on regularized Neural Network and Rough set”

- Proceedings. 2002 International Conference on Machine Learning and Cybernetics, 2002, volume 1, 4-5 Nov. 2002, pages: 315-318.
- [7] L. Yu, W. Shouyang and K.K. Lai, "A Rough-Set-Refined Text Mining Approach for Crude Oil Market Tendency Forecasting". International Journal of Knowledge and Systems Sciences, volume 2, n° 1, March 2005.
- [8] F.E.H. Tay and S. Lixiang, "Economic and Financial prediction using Rough Sets Theory Model". European Journal of Operacional Research. 2002.
- [9] Z. Pawlak, "Why Rough Sets?", Fuzzy Systems. Proceedings of the Fifth IEEE International Conference on, September, 8-11, 1996, vol. 2, pp. 738-743.
- [10] Z. Pawlak, "Rough Sets and data analysis". Fuzzy Systems Symposium, 1996. 'Soft Computing in Intelligent Systems and Information Processing', Proceedings of the 1996 Asian 11-14 Dec. 1996 pages: 1-6.
- [11] S. Haykin, "Neural Networks": A Comprehensive Foundation, N.J., Willey&Sons, 1994.
- [12] A.P. Braga, T.B. Ludermir and A .C.P.L.F. Carvalho, "Redes Neurais Artificiais: Teoria e Aplicações", Rio de Janeiro. Ed. LTC, 2000.
- [13] V. Dhar and R. Stein, "Seven Methods for transforming Corporate Data into Business Intelligence". Upper Saddle River, NJ. Prentice Hall. 1997.
- [14] R. Duda, P. Hart and D. Stork, "Pattern Classification and Scene Analysis", John Wiley Profession, NY, 2000.
- [15] T. Kohonem, "Clustering Taxonomy and Topological Maps of Patterns", Proc. Of the Sixth Intern. Conference on Pattern Recognition, 1982, pp. 114-128.
- [16] Z. Pawlak, J. Grzymala-Busse, R. Slowinski, W. Ziarko. "Rough Sets", Communications of the ACM, pages: 89-95, 1995.
- [17] Available in <ftp://ftp.ics.uci.edu/pub/machine-learning-databases>.
- [18] Available in <http://www.cis.hut.fi/projects/somtoolbox/documentation/start.shtml>.
- [19] Available in <http://www.idi.ntnu.no/~aleks/rosetta>.
- [20] P. Lingras, "Rough Neural Networks". Sixth International Conferences Information Processing and Management of Uncertainly in Knowledge-Based Systems, Proceedings (IPMU96), volume II, july 1-5. Grenada, pages: 1445-1450, 1996.
- [21] P. Lingras, "Unsupervised learning using Rough Kohonem Neural Networks Classifiers", Proc. Of Symp. On Modeling, Analysis and Simulation, CESA 96. Lille France, pages: 753-757, 1996.