COMPUTING THE SIGNIFICANCE OF AN INDEPENDENT VARIABLE USING ROUGH SET THEORY AND NEURAL NETWORK

Renu Vashist*

M.L. Garg**

ABSTRACT

Artificial Neural Network (ANN) and Rough Set Theory (RST) has evolved as a new decision making tool and has been widely applied to a variety of application problem involving classification. ANN and RST are both data mining tool. Neural network can be used to extract knowledge from trained neural networks so that the users can gain a better understanding of the solution, whereas rough set theory can be used to extract knowledge from the information system. The knowledge from the neural network and rough set theory can be represented in the form of rules. The RST and ANN can be used as a tool for reducing and choosing the most relevant sets of attributes from the data set.. In this paper we find the most significant attributes using the rough set theory and neural network and our finding leads to the conclusion that both these methodologies has the same result for a particular dataset.

Keywords: *Artificial Neural Network; Rough set ; Decision Table ; Significant attribute; Reduct ; Core ; Decision making tool; independent Variable.*

*Research Scholar, School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra, (J & K), India.

**Professor and Dean, School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra (J & K), India

1. INTRODUCTION

Z. Pawlak[11] proposed that rough set theory is a comparatively new mathematical instrument to deal with vague and uncertain information. With the development of artificial intelligence and cognitive sciences rough set theory can be used to discover dependency among the attributes, attribute reduction, finding the most significant attribute of information system, generating rules etc[11-12]. On the other hand neural network is a kind of network system that simulates the human brain information processing system. However the processing that takes place in human brain is far more complex. As the artificial neural network now a days becomes a powerful tool to solve problem, because of its strong fault tolerance, self-organization, massively parallel processing, self-adapted and so on, which plays an important in breaking bottleneck of science and technology and exploring the complex phenomenon of nonlinear statistical modeling[4,6].

The neural network and rough sets methodologies have their place among intelligent classification and decision support systems. Both these techniques can be used for finding the significance of the attributes of a dataset but they have their own advantages and disadvantages. The rough set is a powerful tool to process uncertain or high-dimensional data, but it to sensitive to noise and it generates too many rules. This technique also help us to find the relationship among the independent and dependent variable and degree of dependency among the two variables. Whereas neural network has good robustness and self-learning, ability to detect complex nonlinear relationships between dependent and independent variables but for massive data it cannot get good effect, it require massive time frequency. Neural networks often predict with higher accuracy than other techniques because of the networks' capability to fit any continuous functions [2,7,8]. One major drawback often associated with neural networks is their lack of explanation power. It is difficult to explain how the networks arrive at their solutions due to the complex nonlinear mapping of the input data by the networks. The outstanding feature of RST is its simplicity that makes this approach more superior.

2. ARTIFICIAL NEURAL NETWORK

Artificial neural networks are massively parallel adaptive network of simple nonlinear computing elements called neurons which are intended to abstract and model some of the functionality of the human nervous system in an attempt to partially capture some of its computational strengths[1]. A neural network is a massively parallel distributed processor

that has a natural propensity for storing exponential knowledge and making it available for use. It resembles the brain in two respects:

a) Knowledge is acquired by the network through a learning process.

b) Interneuron connection strengths known as synaptic weights are used to store the knowledge[3].

2.1 Architecture of Neural Network

In neural network artificial neurons connects in fundamentally two different architectures, Feedforward and Feedback. In Feedforward network the output depends only on the presently applied input. These networks are static in nature. The basic architecture consists of three types of neuron layers: input, hidden, and output. In feed-forward networks, the signal flows from input to output units, strictly in a feed-forward direction. The data processing can extend over multiple layers of units, but no feedback connections are present[5]. For feed-forward networks, the dynamical properties of the network are important. In the feed back network , the signal flows from output units back to the input units.

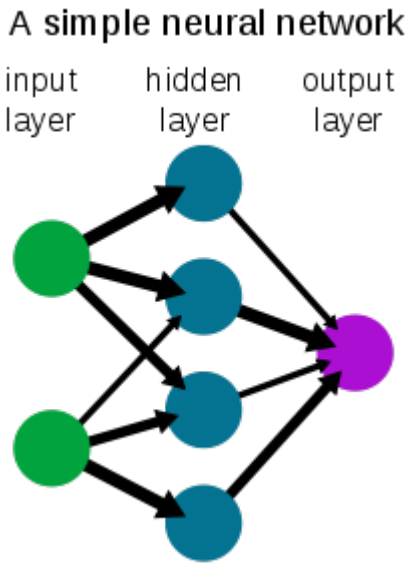


Fig 1 Feed Forward Artificial Neural Network

2.2 Applying Neural Network to Monk's Dataset

The Monk's problem rely on the artificial robot domain, in which robots are described by six different attributes. The data sets were provided by *UCI Repository of machine learning databases* (http://archive.ics.uci.edu/ml/). This dataset is suitable for finding the significance of attribute using Neural Network and Rough Set theory.

For Monk's dataset we have following attributes

Definition of attributes

Variable	Definition
A1	Head_ Shape
A2	Body_Shape
A3	Is_Smiling
A4	Holding
A5	Jacket_Colour
A6	Has_Tie

The learning task is a binary classification task. Each problem is given by a logical description of class. Robot either belong to the class or not. Total of 432 robots are given for classification task. The Artificial Neural Network analysis is done using SPSS 17 in the Window 2000 environment.

Table 1 Case Processing Summary

		N	Percent
Sample	Training	312	72.2%
	Testing	120	27.8%
Valid		432	100.0%
Excluded		0	
Total		432	

The case processing summary shows that 312 cases were assigned to the training sample and 120 to the testing sample. Total number of robots are 432.

Table 2 Network Information

Input Layer	Factors	1	Head Shape
		2	Body Shape
		3	Is Smiling
		4	Holding
		5	Jacket Colour
		6	Has Tie
		Number of Units[a]	17
Hidden Layer(s)		Number of Hidden Layers	1
		Number of Units in Hidden Layer 1[a]	5
		Activation Function	Hyperbolic tangent
Output Layer	Dependent Variables	1	D
	Number of Units		2
	Activation Function		Softmax
	Error Function		Cross-entropy

Table 2 Network Information

Input Layer	Factors	1	Head Shape
		2	Body Shape
		3	Is Smiling
		4	Holding
		5	Jacket Colour
		6	Has Tie
		Number of Units[a]	17
Hidden Layer(s)		Number of Hidden Layers	1
		Number of Units in Hidden Layer 1[a]	5
		Activation Function	Hyperbolic tangent
Output Layer	Dependent Variables	1	D
	Number of Units		2
	Activation Function		Softmax
	Error Function		Cross-entropy

The network information table displays information about the neural network and is useful for ensuring that the specifications are correct. In this table

The **input layer** contains the predictors.

The **hidden layer** contains unobservable nodes, or units. The value of each hidden unit is some function of the predictors, the exact form of the function depends in part upon the network type and in part upon user-controllable specifications.

The **output layer** contains the responses. Each output unit is some function of the hidden units.

Table3 Classification Table with dependent variable D

		Predicted		
Sample	Observed	0	1	Percent Correct
Training	0	141	0	100.0%
	1	0	171	100.0%
	Overall Percent	45.2%	54.8%	100.0%
Testing	0	63	0	100.0%
	1	0	57	100.0%
	Overall Percent	52.5%	47.5%	100.0%

The classification table i.e table3 shows the practical results of using the network. For each case, the predicted response is *Yes* if that cases predicted pseudo-probability is greater than 0.5. For each sample:

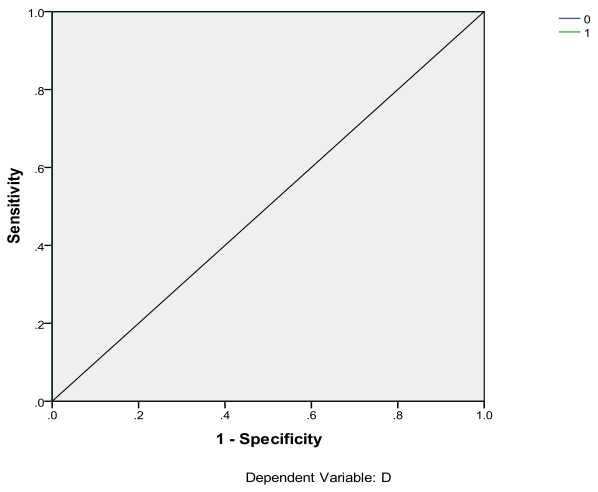
1 Cells on the diagonal of the cross-classification of cases are correct predictions.

2 Cells off the diagonal of the cross-classification of cases are incorrect predictions.

Overall, 100% of the training cases are classified correctly. Here 100% of the cases were correctly classified by the model. This suggests that, overall, our model is in fact a perfect model.

2.3 ROC curve

Graph1 displays an ROC (Receiver Operating Characteristic) curve for each categorical dependent variable. It also displays table4 giving the area under each curve. For a given dependent variable, the ROC chart displays one curve for each category. Since our dependent variable has two categories, then each curve treats the category at issue as the positive state versus the other category. The ROC curve gives you a visual display of the **sensitivity** and **specificity** for all possible cutoffs in a single plot, which is much cleaner and more powerful than a series of tables. The chart shown here displays two curves, one for the category *No* and one for the category *Yes*. This chart is based on the combined training and testing samples. The area under the curve is a numerical summary of the ROC curve, and the values in the table4 represent, for each category, the probability that the predicted pseudo-probability of being in that category is higher for a randomly chosen case in that category than for a randomly chosen case not in that category. For example, for a randomly selected robot that belong to the class and randomly selected robots that does not belong to the class, there is a 100% probability that the model-predicted pseudo-probability of robot belonging to the class and robots that does not belong to the class will be the same. While the area under the curve is a useful one-statistic summary of the accuracy of the network, you need to be able to choose a specific criterion by which robots are classified.



Graph 1

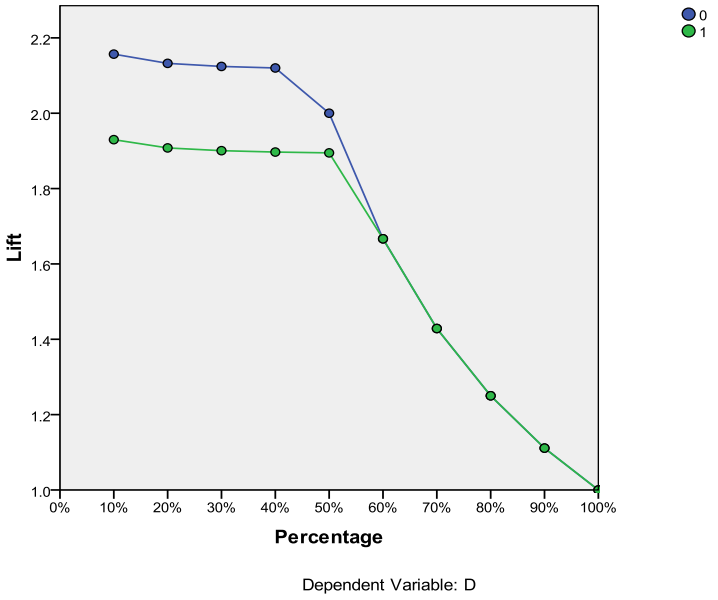
Area Under the Curve

Table 4

		Area
D	0	1.000
	1	1.000

2.4 Lift chart

Graph 2 displays a lift chart for each categorical dependent variable. The display of one curve for each dependent variable category is the same as for ROC curves. The lift chart is derived from the cumulative gains chart; the values on the *y* axis correspond to the ratio of the cumulative gain for each curve to the baseline. Thus, the lift at 10% for the category Yes is 1.90%/10% = 1.9.

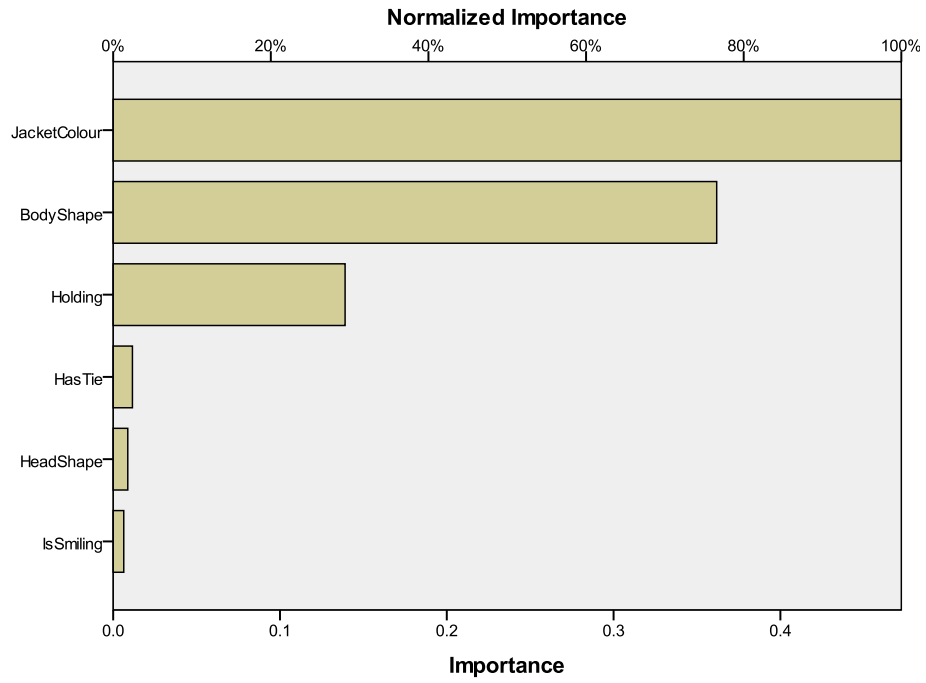


Graph 2

2.5 Independent Variable Importance Analysis

This analysis performs a sensitivity analysis, which computes the importance of each predictor in determining the neural network. The analysis is based on the combined training and testing samples or only on the training sample if there is no testing sample. This creates table5 and a graph3 displaying importance and normalized importance for each predictor. Note that sensitivity analysis is computationally expensive and time-consuming if there are large numbers of predictors or cases. The importance of an independent variable is a measure of how much the network's model-predicted value changes for different values of the

independent variable. Normalized importance is simply the importance values divided by the largest importance values and expressed as percentages.



Graph 3

Table5 Independent Variable Importance

	Importance	Normalized Importance
Head Shape	.009	1.9%
Body Shape	.362	76.6%
Is Smiling	.006	1.3%
Holding	.139	29.4%
Jacket Colour	.473	100.0%
Has Tie	.012	2.4%

Table 5 shows that we have only three significant attribute that is Jacket Colour, Body shape and holding.

3 ROUGH SET THEORY

Rough set theory proposed by Z. Pawlak[10] is a new technique of decision making in the presence of vagueness and uncertainty. Rough set concept can be defined quite generally by means of topological operations, *interior* and *closure*, called *approximations.* Rough set data analysis is used in many applications such as process control, economics, medical diagnosis,

biochemistry, environmental science, biology, chemistry psychology, conflict analysis and other fields can be found in [13-15].

The rough set philosophy is founded on the assumption that with every object of the universe of discourse we associate some information (data, knowledge). The starting point of rough set theory is the indiscernibility relation, generated by information about objects of interest. Objects characterized by the same information are indiscernible (similar) in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of rough set theory [9]. Any set of all indiscernible (similar) objects is called an elementary set and forms a basic granule of knowledge about the universe. The basic assumption of rough set theory as put forth by Pawlak is that human knowledge about a universe depends upon their capability to classify its objects. Equivalence relation is used to define the rough set[11]. Every vague concepts, in contrast to precise concepts, cannot be characterized in terms of information about their elements. Therefore any vague concept is characterized by pair of precise concepts, called the lower and upper approximations of the set, represents a rough set. The lower approximation of a rough set comprises of those elements of the universe, which surely belong to the set with the available knowledge. The upper approximation on the other hand comprises of those elements, which are possibly in the set. The difference between the upper and the lower approximation constitute the boundary region of the vague concept. Boundary region will consist of those objects which we cannot decisively classify into the set [12]. The concept of rough sets was primarily concerned with the study of intelligent systems characterized by insufficient and incomplete information [22].

The basic concept of rough set theory is to deal with the information table or decision table. Every decision table has two set of attributes. One is called the condition attributes and other is called the decision attributes. Basic problems which can be solved using the rough set approach are the following[16-17]:

1) description of objects in terms of attribute values.
2) dependencies (full or partial) between attributes.
3) reduction of attributes.
4) significance of attributes.
5) decision rules generation

Now, we define rough set mathematically.

Let a finite set of objects *U* and a binary relation $R \subseteq U \times U$ be given. The sets *U*, *R* are called the *universe* and an *indiscernibility relation*, respectively. The discernibility relation represents our lack of knowledge about elements of *U*. For simplicity, we assume that *R* is an equivalence relation. A pair (*U*, *R*) is called an *approximation space*, where *U* is the universe and *R* is an equivalence relation on *U*. Let *X* be a subset of *U*, i.e. $X \subseteq U$. Our goal is to characterize the set *X* with respect to *R* [19].

• The set of all objects which can be with *certainty* classified as members of *X* with respect to *R* is called the *R-lower approximation* of a set *X* with respect to *R*, and denoted by

$$\underline{R(X)} = \{x \in U : R(x) \subseteq X\}$$

• The set of all objects which can be only classified as *possible* members of *X* with respect to *R* is called the *R-upper approximation* of a set *X* with respect to *R*, and denoted by

$$\overline{R(X)} = \{x \in U : R(x) \cap X \neq \phi \}$$

• The set of all objects which can be decisively classified neither as members of *X* nor as members of - *X* with respect to *R* is called the *boundary region* of a set *X* with respect to *R*, and denoted by *RN* (*X*) *R* , i.e.

$$RN(X)\,R = \overline{R(X)} - \underline{R(X)}$$

• A set *X* is called *crisp* (*exact*) with respect to *R* if and only if the boundary region of *X* is empty.

• A set *X* is called *rough* (*inexact*) with respect to *R* if and only if the boundary region of *X* is nonempty.

3.1 Information System or Decision Table

The dataset in case of rough set theory is always define in the form of table that is known as information system. An information system or Decision Table can be viewed as a table, consisting of objects(rows) and attributes (column).The attribute set consist of condition and decision attribute. Condition attributes are independent variable and decision attribute is a dependent variable [10]. In our decision table we have six condition attributes i.e Head shape, Body shape, Jacket colour, Holding, Has tie, is Smiling and one decision attribute(Class attribute).

Now we apply rough set theory to Monk's dataset using ROSE 2 Software which was created at the Laboratory of Intelligent Decision Support systems of the Institute of Computing Science in Poznan [14].

We find the Lower and Upper Approximation of Monk's dataset as follows:

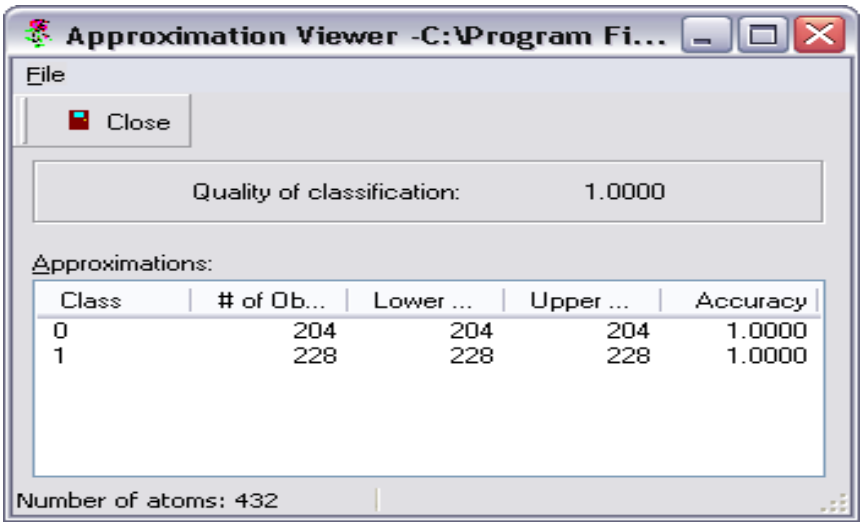


Fig 1

Class	# of objects	Lower Approximation	Upper Approximation	Accuracy
0	204	204	204	1.00
1	228	228	228	1.00

3.2 Dependency of Attributes

Important issue in rough set data analysis is discovering *dependencies* between condition and decision attributes. Intuitively, a set of attributes *D depends totally* on a set of attributes *C*, denoted $C \Rightarrow D$, if all values of attributes from *D* are uniquely determined by values of attributes from *C*. In other words, *D* depends totally on *C*, if there exists a functional dependency between values of *D* and *C*. Formally dependency can be defined in the following way. Let *D* and *C* be subsets of *A*. We will say that *D depends on C* in a *degree k* $(0 \leq k \leq 1)$, denoted $C \Rightarrow_k D$, if $k = \gamma(C, D)$.

If $k = 1$ we say that *D depends totally* on *C*, and if $k < 1$, we say that *D depends partially* (in a *degree k*) on *C*.The coefficient *k* expresses the ratio of all elements of the universe, which can be properly classified to blocks of the partition *U*/*D*, employing attributes *C* [18].Thus the concept of dependency of attributes is strictly connected with that of consistency of the decision table.

If *D depends in degree k*, $0 \leq k \leq 1$, on *C*, then

$\gamma(C,D) = |POSc(D)| / |U|$

Where *POSc(D),* is the positive region of of the partition *U/D* with respect to *C*, it is also called as the lower approximation of the set, it is the set of all elements of *U* that can be uniquely classified to blocks of the partition *U/D*, by means of *C.*

In monk's dataset we have 204 elements in lower approximation for class0 and 228 elements in lower approximation of class 1 and the total element in lower approximation is 432 then the dependency coefficient is calculated as

$\gamma(C,D) = 432/432 = 1$

Here in this cases if $k = 1$, we say that *D depends totally* (in a *degree k*) on *C.*

3.3 Reduction of Attributes Reduct and Core

A fundamental problem that arise in rough set theory is weather the whole set of knowledge base is necessary to define some categories available in the knowledge base. Knowledge dependency play an important role in reduction of knowledge. There are two fundamental concepts, a reduct and the core. A reduct of knowledge essentially reduces the knowledge without compromising the knowledge base. From Reduct we can also generate original knowledge base. In short, reduct is essential part of knowledge, if we loose data from reduct we cannot reproduce the original knowledge, whereas the core is the indispensable part of the knowledge [21-22]. .

Now we define a notion of a core of attributes. Let B be a subset of A. The core of B is a set of all indispensable attributes of B. The following is an important property, connecting the notion of the core and reducts

Core(B) = ∩ Red(B),

where Red(B) is the set off all reducts of B.

Because the core is the intersection of all reducts, it is included in every reduct, i.e., each element of the core belongs to some reduct. Thus, in a sense, the core is the most important subset of attributes, we cannot remove any for of its elements without affecting the classification power of attributes. Finding all the reduct of a dataset is NP hard problem [10]. However, in many applications we do not need to compute all reducts, but only some of them, satisfying specific requirements.

In Monk's dataset we have three attributes in core fig2 and only one reduct fig3.

Core1 = Body Shape

Core 2 = Holding

Core 3 = Jacket Colour

Reduct = {Body shape, Holding, Jacket Colour}

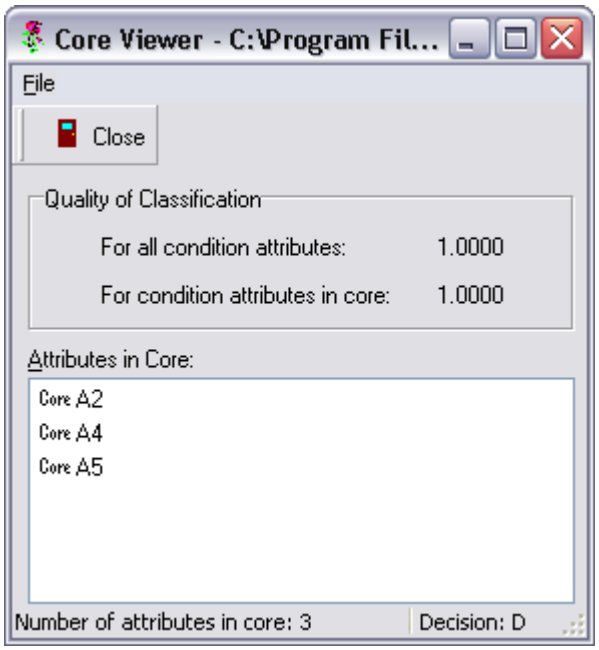


Fig 2

The attributes which are in core are also the most significant attributes and are indispensable. We cannot remove any of the core attribute without affecting the classification task. The attribute which are in core is same as the attribute that we find using the artificial neural network. Both these methodologies find that the attribute A2 i.e Body Shape, A4 that is Holding and A5 that is Jacket Colour are most significant attributes.

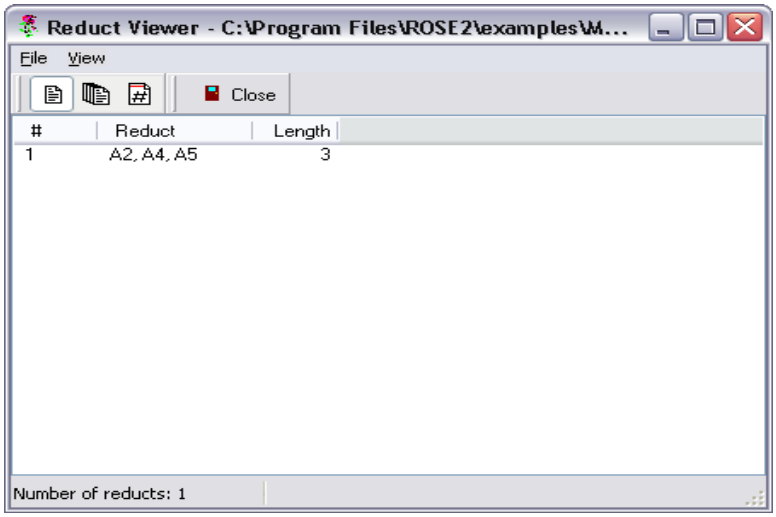


Fig 3

4 CONCLUSION

ANN and RST are relatively new decision making tools used for classification. These techniques have wide range of application in finance, bioinformatics, medicine, economics etc. The central concept in these technique is to find the significance of independent attributes or variables. In this paper we have used both these methodologies for finding the

significance of attribute of a dataset. We concluded that the significant attribute found by both these techniques are same.

REFERENCE

[1] kohonen, T., 'An introduction to neural computing' , *Neural Networks*' **1** (1988) 3-16.

[2] Coakley, J.R., and Brown, C.E. (1993). Artificial neural networks applied to ratio analysis in the analytical review process. *Intelligent Systems in Accounting, Finance and Management, 2*, 19-39.

[3] Desai, V.S., and Bharati, R. (1998). The efficacy of neural networks in predicting returns on stock and bond indices. *Decision Sciences*, *29*(2), 527-544.

[4] Dutta, S., Shekhar, S., and Wong, W.Y. (1994). Decision support in non—conservative domains: Generalization with neural networks. *Decision Support Systems*, *11*(5), 527-544.

[5] Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*, 251-257.

[6] Salchenberger, L.M., Cinar, E.M., and Lash, N.A. (1992). Neural networks: A new tool for predicting thrift failures. *Decision Sciences*, *23*(4), 899-916.

[7] Tam, K.Y., and Kiang, M.Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science*, *38*(7), 926-948.

[8] Tana, S.S., and Koh, H.C. (1992). A multi-layer perceptron model of credit scoring for assessing default risk in charge card applicants. *International Journal of Management*, *14*(2), 250-255.

[9] Krusinska E., Slowinski R., Stefanowski J.: Discriminant Versus Rough Set Approach to Vague Data Analysis. *Applied Stochastic Models and Data Analysis* 8 (1992) 43-56..

[10] Pawlak Z.: *Rough sets. International Journal of Computer and Information Sciences* 11, (1982) 341-356.

[11] Pawlak Z.: *Rough Sets - Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht, Boston, London, 1991.

[12] Pawlak Z., Skowron A.: Rough Membership Functions. In: Yaeger R.R., Fedrizzi M. and Kacprzyk J., eds. *Advances in the Dempster Shafer Theory of Evidence*, John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore (1994), 251-271.

[13] Pawlak Z., Slowinski R.: Rough Set Approach to Multi-Attribute Decision Analysis, Invited Review. *European Journal of Operational Research* **72** (1994) 443-459.

[14] Prędki B, Wilk S (1999) Rough set based data exploration using ROSE system. In: Ras ZW, Skowron A (Eds.), Foundations of Intelligent, Lecture Notes in Artificial Intelligence, vol. 1609, Springer, Berlin, 172-180

[15] Skowron A., Rauszer, C.: The Discernibility Matrices and Functions in Information Systems. In: Slowinski R. ed. *Intelligent Decision Support – Handbook of Advances and Applications of the Rough Set Theory*. Kluwer Academic Publishers, Dordrecht, Boston, London (1992), 311-362.

[16] Slowinski R. ed.: *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory*. Kluwer Academic Publishers, Dordrecht, Boston, London, 1992.

[17] Slowinski R., Stefanowski J.: Rough classification with valued closeness relation.In: E.Diday, Y.Lechevallier, M.Schrader, P.Bertrand, B.Burtschy (eds.), *New Approaches in Classification and Data Analysis*. Springer-Verlag, Berlin, 1994, pp. 482-489.

[18] Ziarko W. ed.: *Rough Sets, Fuzzy Sets and Knowledge Discovery. Proceedings of the International Workshop on Rough Sets and Knowledge Discovery* (*RSKD'93*). Banff, Alberta, Canada, 12-15 October 1993, Springer Verlag, 1993.

[19] Z. Pawlak, Information systems – theoretical foundations, Information Systems 6 (1981) 205–218..

[20] Z. Pawlak, Rough classification, International Journal of Man-Machine Studies 20 (5) (1984) 469–483.

[21] Z. Pawlak, Rough logic, Bulletin of the Polish Academy of Sciences, Technical Sciences 35 (5–6) (1987) 253–258. [17] Polkowski, L. & Skowron, A. (Eds.) (1998c). Rough sets in knowledge discovery, Vol. 2. Heidelberg: Physica–Verlag.