Missing data estimation on heart disease using Artificial Neural Network and Rough Set Theory

Conference Paper · December 2007

DOI: 10.1109/ICIAS.2007.4658361 · Source: IEEE Xplore

3 authors, including:

Noor Akhmad Setiawan
Gadjah Mada University
63 PUBLICATIONS **153** CITATIONS

SEE PROFILE

Ahmad Fadzil Mohd Hani
Universiti Teknologi PETRONAS
147 PUBLICATIONS **846** CITATIONS

SEE PROFILE

Missing Data Estimation on Heart Disease Using Artificial Neural Network and Rough Set Theory

[1]N.A. Setiawan, [2]P.A. Venkatachalam, [3]A.F.M. Hani
Electrical and Electronic Engineering Programme,
Universiti Teknologi PETRONAS
Bandar Seri Iskandar 31750 Tronoh, Perak, MALAYSIA
Email : [1]noor_akhmad@utp.edu.my, [2]paruvachiammasai_venkatachala@petronas.com, [3]fadzmo@petronas.com.my

***Abstract*-The objective of this research is to implement a method for estimating the real missing data in heart disease datasets and to show how it affects the resulting knowledge. Missing data is common problem in Knowledge Discovery from Database (KDD) processes that can lead significant error in extracted knowledge. We use hybridization of Artificial Neural Network and Rough Set Theory (ANNRST) to estimate the real missing data on heart disease from UCI (University of California, Irvine) datasets [1]. ANN with reduced input features is used to estimate the missing data. RST is used to reduce the dimensionality of input features and to extract the knowledge as reducts and rules from heart disease datasets with estimated missing data. RST, decomposition tree, Local Transfer Function Classifier (LTF-C) and k-Nearest Neighbor (k-NN) classifier are used to calculate the accuracy. Comparative study with k-NN estimation, most common attribute value filling and deletion of missing data are made to evaluate the extracted knowledge. ANNRST can be considered as the appropriate estimation method when strong relationship between original complete datasets and estimated datasets is important (the estimated datasets really represent the nature of original complete datasets) as it gives the best accuracy and coverage for almost all the classifiers.**

Keywords: neural network, rough set theory, missing value.

I. INTRODUCTION

Knowledge discovery from data (KDD) processes usually encounter missing data problem. The source of problem may be from collecting processes due to the device and human error. The quality of data is very important factor for KDD to reach good quality [1].

Many methods to deal with missing data have been proposed. Grzysmala-Busse and Hu [2] compared several approaches to deal with missing data in data mining. The comparison on approaches to assign missing attribute values was also studied by Li and Cercone [3]. A new approach, RSFit and ItemRSFit on processing data with missing attribute values based on rough set theory (RST), distance based methods and association rules were proposed [4,5]. Al Shalabi, Najjar and Al Kayed proposed a framework to deal with missing data by evaluating reducts and rules generated by RST with four different methods of imputation [6]. Researches on missing data imputation have also been carried in specific fields and problems. Bhattacharya, Shrestha and Solomatine [7] used ANN in reconstructing missing wave data in sedimentation modeling, and found that ANN gives the reasonable accuracy. ANN has also been used to predict missing data of wind speeds [8]. Comparison between the methods for missing data prediction (data imputation): univariate methods (linear, spline and nearest neighbor interpolation), multivariate (Regression-based imputation, Nearest Neighbor (NN), Self-Organizing Map (SOM), Multi-Layer Perceptron (MLP)), and hybrid methods of the above by using simulated annealing missing data patterns on air quality datasets were used [9]. Three methods for missing value estimation for DNA microarrays: a Singular Value Decomposition (SVD) based method (SVDImpute), weighted k-nearest neighbors (KNNImpute) and row average were implemented and evaluated [10]. KNNImpute appears to provide more robust and sensitive method than SVDImpute and row average.

The performance of ANN depends on the training dataset. The high dimensionality of input attributes from training dataset may degrade the learning performance of ANN [11]. High dimensionality can lead poor generalization and convergence difficulties during training process and needs more computational resources and memory. Attribute reduction is required to get over the problem of dimensionality curse. Reduct concept of rough set theory can be used to reduce the dimensionality of attributes without loss of information [12]. Setiawan, et al. [1] proposed ANNRST to predict simulated missing value of heart disease dataset. It is shown that ANNRST has comparable maximum accuracy and better average accuracy than ANN method when it is used to predict simulated missing value. It also compares the proposed method with k-NN imputation, ANN with Piecewise Linear Network with Orthonormal Least Square (PLN-OLS) feature selection and most common attribute value filling. It has been shown that ANNRST works better.

ANN with RST attribute reduction as in [1] is implemented in this research to predict the real missing values on heart disease data from UCI database. The evaluation of generated reducts, rules and accuracy after estimation of missing data using ANNRST is compared with k-NN imputation, Most Common attribute value Filing (MCF), Concept Most Common attribute value filling (CMCF) which is most common attribute value filling according to its decision attributes and Deletion of Missing values (DM).

1-4244-1355-9/07/$25.00 @2007 IEEE

II. Backpropagation Neural Network

Back-propagation or Multi-Layer Perceptron (MLP) is the most common ANN used in many research areas. It is composed of connected nodes (neurons) and weights as processing elements. Fig. 1 shows single hidden layer MLP, where circles represent nodes and arrows represent the forward propagated function signals that have their respective weights. The knowledge is represented as weights between the layers represented as w_{ij} in neuron input-output relationship of (1).

$$y_j^{(l)} = \Phi(v_j^{(l)}) = \Phi\left(\sum_{i=0}^{p} w_{ij}^{(l)} x_{ij}^{(l)}\right) \quad (1)$$

where l denotes the layer number, $y_j^{(l)}$ denotes the output of the jth neuron in the lth layer, $v_j^{(l)}$ denotes the weighted sum of neuron's input, $x_{ij}^{(l)}$ denotes the ith input of the neuron, $w_{ij}^{(l)}$ denotes the contribution weights of the ith input to the neuron, and $\Phi(\bullet)$ is the activation function of the neuron [13]. The activation function is a smooth nonlinear function. It can have several forms, such as logistic function and hyperbolic tangent function. Back propagation and its variant is the most common learning algorithm in MLP. It consists of two steps, forward direction and backward direction. In the backward direction, weights are updated according to the error minimization process.

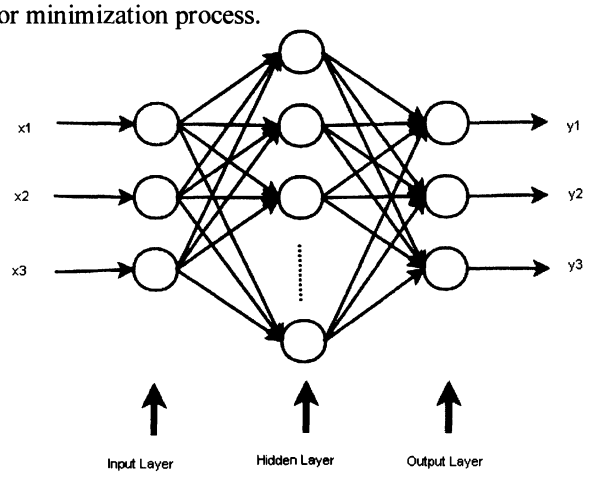

Fig. 1. Topology of a single hidden layer MLP [1].

In the present work, three layered network with eight nodes of hidden layer and single node output layer are used. Number of input nodes may be used based on results of feature selection and attribute reduction by RST.

III. Rough Set Theory

Rough set theory (RST) deals with the analysis of classification of a set of objects that may represent vagueness of knowledge. RST allows the way to discern and classify objects in data sets of this type, when specifying the objects into certain categories is impossible [14, 15]. Consider an example of data set shown in Table I.

TABLE I
EXAMPLE OF DATASET

$x \in U$	a	b	c	$\Rightarrow$	d
0	1	1	2		1
1	2	1	1		2
2	1	0	0		1
3	0	1	1		2
4	1	0	2		1
5	2	2	0		1

The data set can be represented as decision table defined as:

$$S = (U, A, d) \quad (2)$$

where U is the set of objects, A is the set of *conditional attributes* and d represents the *decision attributes*.

Let $B \subseteq A$. Then each subset defines an equivalence relation called an *indiscernibility relation* ($IND_A(B)$) which is defined as:

$$IND_A(B) = \{(x, y) \in U^2 \mid \forall a \in B, a(x) = a(y)\}. \quad (3)$$

For example for B = $\{c\}$, equation (3) will induce a partition of U into sets that use only attributes in B, each object in a set cannot be distinguished from other objects in the set. Thus the partition of $IND_A(B)$ is denoted as $U/IND_A(B)$ $=\{\{0,4\},\{1,3\},\{2,5\}\}$. The sets which the objects are divided into are called *equivalence classes* denoted as $[x]_B$.

Concept of *set approximation* is needed in order to classify an object based only on the equivalence class in which it belongs. Let $B \subseteq A$. The approximation of a set of objects, X, using only the information in B is defined as:

B-lower approximation of X:

$$\underline{B}X = \{x \mid [x]_B \subseteq X\}. \quad (4)$$

B-upper approximation of X:

$$\overline{B}X = \{x \mid [x]_B \cap X \neq \varnothing\}. \quad (5)$$

The lower approximation is the set consisting of all objects for which the object's equivalence class is a subset of the approximated set. This set contains all objects which with certainly belong to set X.

The upper approximation is the set consist of all objects for which the intersection of the object's equivalence class and the approximated set is not an empty set. This set contains all objects which possibly belong to the set X.

B-boundary Region of X:

$$BR_B(X) = \overline{B}X - \underline{B}X. \quad (6)$$

This set contains the objects that can neither be classified as definitely inside X nor definitely outside X. A set is *rough* if $BR_B(X) \neq \varnothing$. For B= $\{c\}$, with concept X $=\{1,2,3\}$ then $\underline{B}X$ $=\{1,3\}$ and $\overline{B}X$ $=\{1,2,3,5\}$ and $BR_B(X)$ $=\{2,5\}$. Consider d as decision attribute. Then *B-positive region* of d is defined as:

$$POS_B(d) = \cup_{X \in U/d} \underline{B}X. \quad (7)$$

where, *U/d* represents the partition of *U* according to decision class *d*. In Table 1, with $B = \{c\}$

$$POS_B(d) = \cup\{\{0,2,4,5\},\{1,3\}\} = \{0,1,2,3,4,5\}.$$

Reducing the knowledge results in *reducts*. A reduct is a minimal set of attributes of $B \subset A$, such that

$$POS_B(d) = POS_A(d). \quad (8)$$

A reduct is combination of attributes that able to discern between objects as well as all attributes. Reducts can be computed based on *discernibility matrices* and *discernibility function*. A discernibility matrix of decision system *S* is a symmetric $n \times n$ matrix with entries:

$$c_{ij} = a \in A \mid a(x_i) \neq a(x_j) \text{ and } d(x_i) \neq d(x_j) \quad (9)$$

for $i, j = 1,...,n$.

The entries of each object are thus the attributes that are needed in order to discern object *i* from object *j* relative to the decision. Example of decision relative discernibility matrix is shown in Table II.

Discernibility function can be built from discernibility matrix. A discernibility function f_A for a decision *S* is a Boolean function of *m* Boolean variables $a_1^*,...,a_m^*$ (corresponding to the attributes $a_1,...,a_m$) defined as

$$f_A(a_1^*,...,a_m^*) = \wedge\{\vee c_{ij}^* \mid 1 \le j \le i \le n, c_{ij} \neq \varnothing\} \quad (10)$$

where $c_{ij}^* = \{a^* \mid a \in c_{ij}\}$. By finding the set of all prime implicants of the discernibility function, all the minimal reducts of the system may be determined.

For example,

$$f_A(a,b,c) = \{a \vee b \vee c\} \wedge \{a \vee c\} \wedge \{b \vee c\}.$$

After simplification

$$f_A(a,b,c) = \{a \vee c\} \wedge \{b \vee c\} = \{c\} \vee \{a \wedge b\}.$$ Thus the reducts are $\{a,b\}$ and $\{c\}$. From the reducts computed from this discernibility matrix, decision rules for classification of the objects can be generated.

TABLE II
DECISION RELATIVE DISCERNIBILITY MATRIX

$x \in U$	0	1	2	3	4	5
0						
1	*a,c*					
2		*a,b,c*				
3	*a,c*		*a,b,c*			
4				*a,b,c*		
5		*b,c*		*a,b,c*		

In this research, the concept of reduct is used to reduce the dimension of input attributes for ANN.

IV. EXPERIMENT AND RESULTS

A. Data and Preprocessing

The source of heart (coronary artery) disease data is from data mining repository at UCI [16]. The attributes are shown in Table III. The database consists of four sources: Cleveland Clinic Foundation US (Cleveland), Hungarian Institute of Cardiology Budapest (Hungarian), V.A. Medical Center Longbeach CA US (Longbeach), and University Hospital Zurich Switzerland (Swiss). The amount of data is 920 instances. Cleveland data is the most complete, although it has six missing data, four on *ca* attribute and two on *thal* attribute. Swiss data is the most incomplete, thus it is not used in this research. Longbeach and Hungarian data have many missing values in three of 13 conditional attributes. Then the missing values of three attributes will be estimated. After removing the instances that have too many missing attribute values, the data consists of 661 instances and has 13 conditional attributes, single decision attribute with 351 instances that has missing values on *slope, ca* and *thal* attributes.

To find the reducts, preprocessing is done by discretisation of dataset. The result of diagnosis of coronary artery disease is the decision attribute. Discretisation is conducted using Boolean reasoning.

B. Experiment

Reduct computation using ROSETTA [17] on complete dataset with Boolean reasoning discretisation and Johnson's algorithm results in reduct with six attributes: *age, trestbps, chol, thalach, oldpeak* and *ca*.

ANNRST is used to predict the ten simulated missing values on *slope, ca* and *thal* attributes. The topology of ANNRST: six input nodes. Attributes used for ANNRST are those of the computed reduct (without *ca)* but with decision attribute (*num*). One hundred simulations with random initial weight are conducted on each simulated missing data using MATLAB with Nguyen-Widrow weight initialization, min-max scaling, resilient backpropagation training, and tansig activation function at hidden and output layer [18]. The best accuracy of 100 simulations is chosen to predict the real missing values on *slope, ca* and *thal* attributes. K-NN imputation method, MCF and CMCF to estimate the missing value are also used. Deletion of rows that have missing values is also carried. It consists of only 297 instances. There are five datasets after missing data estimation (including deletion method).

Reducts and rules are generated using genetic algorithm (GA) methods with object related discernibility that are available in ROSETTA on five datasets on combined original complete dataset and estimated dataset.

There are two methods to find the accuracy. First, all the five datasets are split randomly with splitting factor of 0.5 [17]. Fifty percent of 661 (331) instances are for training purpose and the remaining 330 instances are for testing purpose. The accuracies and coverage of all datasets are calculated using different classifiers (RST, decomposition tree, LTF-C, and k-NN).

Second, all the dataset except deletion method dataset is split into two, the first is Cleveland dataset as testing dataset that has only six missing values and the second is estimated datasets (Hungarian and Longbeach) as training dataset. Then accuracy and coverage of all four datasets are calculated with the same classifier as the first method. This second method is

to find the quality of estimated datasets according to original complete datasets (which is Cleveland dataset). All the accuracy and coverage calculations are done using the software RSES [19].

TABLE III
SUMMARY OF ATTRIBUTES (UCI HEART DISEASE DATABSE)

Attribute	Description	Value description
age	Age	Numerical
sex	Sex	1 if male; 0 if female
cp	Chest pain type	1 typical angina 2 atypical angina 3 non-anginal pain 4 asymptomatic
trestbps	Resting systolic blood pressure on admission to the hospital (mmHg)	Numerical
chol	Serum cholesterol (mg/dl)	Numerical
fbs	Fasting blood sugar over 120 mg/dl ?	1 if yes 0 if no
restecg	Resting electrocardiographic results :	0 normal 1 having ST-T wave abnormality 2 LV hypertrophy
thalach	Maximum heart rate achieved	Numerical
exang	Exercise induced angina?	1 if yes 0 if no
oldpeak	ST depression induced by exercise relative to rest	Numerical
slope	The slope of the peak exercise ST segment	1 upsloping 2 flat 3 downsloping
ca	Number of major vessels colored by fluoroscopy	Numerical
thal	Exercise thallium scintigraphic defects	3 normal 6 fixed defect 7 reversible defect
num	Diagnosis of heart disease (angiographic disease status / presence of coronary artery disease (CAD))	0 if less than 50% diameter narrowing in any major vessel (CAD no) 1 if more than 50% (CAD yes)

C. Results and Discussions

The comparison results of reducts and rules between estimation methods using GA reduct computation with object related discernibility are shown in Table IV. The results show that ANNRST gives more reducts and rules with higher reduct and mean rule length than k-NN estimation, CMCF and DM method except MCF which gives more rules and reducts with higher length than ANNRST. CMCF gives the lower number of reduct with the shortest length.

Table V shows that the accuracy of ANNRST is better than DM and MCF for all classifier accuracy but worse than k-NN and CMCF except decomposition tree. For decomposition tree, DM has better accuracy but worse coverage than ANNRST. CMCF is the best in general than other methods for both accuracy and coverage. Table VI shows that ANNRST is the best method among others except for decomposition tree classifier.

TABLE IV
REDUCT AND RULE COMPARISON OF MISSING DATA ESTIMATION METHODS

Estimation methods	Number of rules	Number of reducts	Rule length mean	Reduct length mean
ANNRST	9020 with 8943 deterministic	3765	5.1338	5.4653
k-NN	7010 with 6996 deterministic	3241	5.0366	5.3542
DM	4807 deterministic	2568	4.8982	5.1016
MCF	9572 with 9406 deterministic	3924	5.2000	5.5527
CMCF	6172 deterministic	1740	3.9053	4.4954

TABLE V
ACCURACY AND COVERAGE COMPARISON OF MISSING DATA ESTIMATION METHODS WITH RANDOMIZED SPLIT DATASETS

Estimation methods	Accuracy (Coverage)			
	RST	Decomposition Tree	LTF-C	k-NN classifier
ANNRST	0.855 (1)	0.858 (0.661)	0.876 (1)	0.87 (1)
k-NN	0.891 (1)	0.891 (1)	0.897 (1)	0.927 (1)
DM	0.804 (1)	0.864 (0.595)	0.473 (1)	0.77 (1)
MCF	0.794 (1)	0.815 (0.673)	0.806 (1)	0.8 (1)
CMCF	0.915 (1)	0.943 (0.797)	0.894 (1)	0.924 (1)

TABLE VI
ACCURACY AND COVERAGE COMPARISON OF MISSING DATA ESTIMATION METHODS BY SPLITTING DATASETS TO COMPLETE AND ESTIMATED DATASETS

Estimation methods	Accuracy (Coverage)			
	RST	Decomposition Tree	LTF-C	k-NN classifier
ANNRST	0.825 (1)	0.799 (0.525)	0.818 (1)	0.828 (1)
k-NN	0.789 (1)	0.799 (0.884)	0.739 (1)	0.769 (1)
MCF	0.779 (1)	0.691 (0.182)	0.65 (1)	0.759 (1)
CMCF	0.805 (1)	0.846 (0.749)	0.703 (1)	0.766 (1)

The results show that real missing data estimation is difficult task. The quality of estimation cannot be measured using single criteria. It depends on the nature of datasets and the task of KDD. The priority of KDD parameter will determine the best method of missing data estimation. Reducts, rules, accuracy, coverage or the strong relationship between original complete datasets and estimated datasets can be used as parameter to choose the best estimation methods. ANNRST is the best choice if the strong relationship between original complete dataset (Cleveland) and estimated missing values on Hungarian and Longbeach datasets is the priority as shown in Table VI. The estimated datasets by ANNRST can classify the original complete dataset better than the other estimation methods. The better accuracy of estimation

method on simulated missing data as in [1] is not always gives the better accuracy of its extracted knowledge when it is applied on real missing attribute values of dataset as shown on Table IV and V for ANNRST. The accuracy is not always the main concern of KDD processes. The importance and interestingness of rules and reducts may be considered to get the more useful knowledge but with the acceptable accuracy and coverage. The simple method likes CMCF can be considered superior on accuracy and coverage on randomized split datasets. It also gives small number and the shortest length of rules and reducts. But it can be argued that CMCF method is only most common value filling method (without learning capability of relationship between attributes) and so it does not really represent the real missing data. ANNRST and k-NN methods learn from the nature and properties of complete datasets to estimate missing values of other datasets. They can be considered to give more knowledge, based on complete datasets than MCF and CMCF even though they give worse evaluation of rules, reducts, accuracy and coverage.

V. Conclusion

In this paper, ANN with RST (ANNRST) attribute reduction is implemented and evaluated to predict the real missing attribute values on heart disease data from UCI database. Comparison has been made with the k-NN estimation method, MCF, CMCF, and DM. Three methods of comparison have been conducted. First, comparing reducts and rules. Second is comparing accuracy and coverage of four classifiers on randomized split datasets. Third, splitting datasets to its completed and estimated values, and then using completed datasets to test the accuracy and coverage of classifier based on estimated datasets. ANNRST gives the best accuracy and coverage in general for the third method. CMCF outperforms ANNRST and the others for second method. For the first method, the result of ANNRST depends on the priority of the task of KDD. ANNRST can be considered as the appropriate estimation method when strong relationship between original complete datasets and estimated datasets is important.

Acknowledgment

I would like to thank Universiti Teknologi PETRONAS for the kind support of presenting this paper at ICIAS 2007.

References

[1] N.A. Setiawan, P.A. Venkatachalam, A.F.M. Hani, "Missing Attribute Value Prediction Based on Artificial Neural Network and Rough Set Theory", unpublished.

[2] J.W. Grzysmala-Busse, M. Hu, "A comparison of several approaches to missing attribute values in data mining", *RSCTC 2000, LNAI*, pp. 378-385, 2005.

[3] J. Li, N. Cercone, *Comparisons on different approaches to assign missing attribute values*, Technical Report, CS-2006-04, School of Computer Science, University of Waterloo, January 2006.

[4] J. Li, N. Cercone,"Assigning missing attribute values based on rough set theory", *IEEE GrC 2006*, Atlanta USA, May 2006.

[5] J. Li, N. Cercone, *Predicting missing attribute values based on Frequent Itemset and RSFit*, Technical Reprot, CS-2006-13, School of Computer Science, University of Waterloo, April 2006.

[6] A. Al Shalabi, M.N. Najjar, A. Al Kayed, "A framework to deal with missing data in data sets", *Journal of Computer Science*, Vol. 2(9), pp. 740-745, 2006.

[7] B. Bhattacharya, D.L. Shrestha, D.P. Solomatine, "Neural networks in reconstructing missing wave data in sedimentation modeling", *Proceedings of the XXXth IAHR Congress*, Greece, August 2003.

[8] P. Siripitayananon, H.C. Chen, K.R. Jin, "Estimating missing data of wind speeds using neural network", Proceedings *IEEE SoutheastCon 2002*.

[9] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, M. Kolehmainen, "Methods for imputation of missing values in air quality data sets", *Atsmospheric Environment*, Vol. 38, pp. 2895-2907, 2004.

[10] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, "Missing value estimation methods for DNA microarrays", *Bioinformatics* Vol. 17, No. 6, pp. 520-525, 2001.

[11] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1994.

[12] Q. Shen, A. Chouchoulas, "Rough set-based dimensionality reduction for supervised and unsupervised learning", *Int. J. Appl. Math. Comput. Sci.*,2001, Vol.11, 583-601.

[13] Z. Wang, *Artificial intelligence applications in the diagnosis of power transformer incipient faults*, PhD thesis, Virginia Polytechnic Institute and State University, 2000.

[14] R. Jensen, *Combining rough and fuzzy sets for feature selection*, PhD thesis, University of Edinburgh, 2005.

[15] T.R. Hvidsten, *Fault diagnosis in rotating machinery using rough set theory and ROSETTA*, Technical Report, Norwegian University of Science and Technology, 1999.

[16] D.J. Newman, S. Hettich, C.L. Blake, C.J. Merz,. *UCI Repository of machine learning databases.* http://www.ics.uci.edu/~mlearn/MLRepository.html. Irvine, CA: University of California, Department of Information and Computer Science. 1998

[17] A. Ohrn., *ROSETTA*, http://www.idi.ntnu.no/~aleks/rosetta, 1999

[18] H. Demuth, M. Beale, M. Hagan, Neural *Network Toolbox User's Guide*, The MathWorks, Inc., 2006.

[19] A. Skowron, et al., *RSES 2.2 User's Guide* http://logic.mimuw.edu.pl/~rses., Warsaw University, Poland, 2005