A Computational Intelligence Technique for Effective and Early Diabetes Detection using Rough Set Theory

Kamadi VSRP Varma
GIT,GITAM University
Rushikonda
Visakhapatnam - India

Allam Apparao
CRRao AIMSCS
UoH Campus
Hyderabad - India

P V Nageswar Rao
GIT,GITAM University
Rushikonda
Visakhapatnam - India

ABSTRACT
Huge amount of medical databases requires sophisticated techniques for storing, accessing, analysis and efficient use of stored acquaintance, knowledge and information. In early days intelligent methods like neural networks, support vector machines, decision trees, fuzzy sets and expert systems are widely used in the medical fields. In recent years rough set theory is used to identify the data associations, reduction of data, data classification and for obtaining association rules form the mined databases. In this research contribution we proposed a method for generating association classification rules for the classification of Pima Indian Diabetes (PID) data set taken from UCIML repository. We obtained promising results with this method on the PID data set.

General Terms
Expert systems, fuzzy sets, decision trees

Keywords
Rough sets, Fuzzy sets, Expert system, Pima Indian Diabetes (PID) data set

1. INTRODUCTION
The 21st century technology made it very easy to handle huge quantity of data as a part of storage and analysis. Trillions of databases are used to handle business management, scientific research, medical filed and engineering data management and many other possible applications. The growing databases throughout all fields insisting the new technologies and tools will be developed where that can involuntarily and cleverly transform the processed data into useful knowledge and information. Knowledge discovery database (KDD) is one of the processes used to transform data into knowledge. KDD is the process of extracting understandable patterns, legitimate, novel and potentially useful from the huge groups of data (1). KDD is has three predefined operational stages that is preprocessing, data mining and the post processing. KDD may be used to identify associations, identification of sequences, identification of patterns, Classification, Regression, Clustering, summarization and forecasting the data. KDD process and data mining are interchangeable in the research field (2).

Rough set theory is a novel mathematical technique to deficient knowledge. The deficient knowledge has become a vital topic for computer engineers, particularly in the area of artificial intelligence and machine learning. The most successful methods to handle this deficient or imperfect knowledge is fuzzy set theory proposed by Zadeh in the year 1965 (3) and rough set theory proposed by Pawlak in 1982 (4). Dariusz G. Mikulski used rough set based splitting criterion for binary decision tree classifiers approach for the classification of real time data sets taken from UCILMR (5).The rough set theory is very popular to solve the problems of AI and cognitive sciences, especially in the areas of machine learning, knowledge acquisition, decision analysis, support systems, expert systems, computational tools, knowledge discovery from huge databases, and pattern identification. The rough set theory is used to handle qualitative data and fits into most real life applications adequately. Rough set can be used in different stages of the knowledge engineering process, as attribute identification, extraction and selection, data reduction, decision rule generation and pattern identification (6). Rough set contains multi-phases like, discretization, reducts and rules generation on training set; and classification on test data set. Rough set offers useful methods that are valid in many branches of Artificial Intelligence methods and significant results have been found. Rough set theory resolved many complex problems that attracted the sight of huge researchers in recent years.

Mrozek and K Cyan exhibited a hybrid technique which uses rough set theory to define the objective function for space search of a feature extractor, and neural network to model the uncertain system of automatic diffraction pattern recognition based on rough set theory and neural network (7). Kostek proposed a prototype system to induce generalized rules that explain the association between acoustical factor of concert halls and sound dispensation algorithms using rough set theory (8). Lambert_Torres et al., used rough set theory for power system operation problems and they achieved prominent results (9). The hierarchical rough set based approach is used for sunspot classification by Nguyen et al., (10). Xie et al., used the combination of fuzzy and rough set theory for the construction of intelligent control systems (11). The author He et al., proposed a hybrid model by combing Ant Colony algorithm and rough set theory for the implementation of feature selection algorithm (12).

Rough set has shed light on many research areas, but infrequently found its way into real world applications like medical and clinical data analysis. It is motivated us to use rough set concepts to develop a computational tool for early diabetes detection. We used Pima Indian Diabetes (PID) data set to exhibit the computational performance of the method.

2. Data description
Knowler et al., studies reported a high prevalence of diabetes among Pima Indian women living at Phoneix, Arizona, USA (13). We used Pima Indian Diabetes Data set which is taken from UCIML warehouse to measure the performance of the present model. PID has eight conditional attributes and two

decisions attributes which are known as class labels or class attributes. We eliminated the missing and disgusted data items from the data set to achieve proper performance. The report of PID data set and brief statistical data is presented in Table 1 and Table 2 respectively.

Table 1. Pima Indian Diabetes Data set description.

Attribute	Variable	Abbrev-ation	Specification	
A1	V1	Pregnant	Number of times pregnant	
A2	V2	Glucose	Plasma glucose concentration at 2 hours in an oral glucose tolerance test - (mg/dl)	
A3	V3	DBP	Diastolic blood pressure-(mm Hg)	
A4	V4	TSFT	Triceps skin fold thickness-(mm)	
A5	V5	INS	2-Hour serum insulin-(mu U/ml)	
A6	V6	BMI	Body mass index-(kg/m2)	
A7	V7	DPF	Diabetes pedigree function	
A8	V8	Age	Age	
CLASS		DM	1	tested_Positive:(diabetic)
			2	tested_Negative:(non diabetic)

Table 2. Brief statistical report of PID dataset

Attribute	Mean	Standard deviation	Min/max
1	3.8	3.4	1/17
2	120.9	32	56/197
3	69.1	19.4	24/110
4	20.5	16	7/52
5	79.8	115.2	15/846
6	32	7.9	18.2/57.3
7	0.5	0.3	0.0850/2.3290
8	33.2	11.8	21/81

3. Proposed model

The empirical data is represented as matrix with individual sample items as rows and attributes as columns.

The general information function of the empirical data is represented with an equation as

$$\rho : O \times M \to N$$

Where

O is set of all objectives in the data set.

M is set of all attributes

N is set of all attribute values.

The set of all attributes (M) are identified as conditional attributes (C) and decision attributes (D). Set of values for an attribute $a \in M$ for categorical case is defined as:

$$N^a = \{n / \forall x \in O : n \in \rho(x, a)\}$$

For non-categorical case with a split threshold λ ,

$$N^a = \begin{Bmatrix} n / n \in \{ture,\ false\} \wedge \forall x \in O \wedge \lambda \in R \\ : n = (\rho(x, a\langle\lambda\rangle)) \end{Bmatrix}$$

Unique values for an attribute $a \in M$

$$ON^a = \{n / n \in (N^a \cap N^a)\}$$

Constructing Indesernibility matrix between the values of attribute $c \in C$ and the values of attribute D.

$e = card(ON)^C$ and $f = card(ON)^C$

Framing the dependency matrix:

$$W_{ij}^C = \begin{Bmatrix} p \in O / INDESERINIBILITY(C \cup D) / c \in \\ C \wedge \forall x \in p : (i, \rho(x, c)) \in \langle ON^c \rangle \\ \wedge (j, \rho(x, D)) \in \langle ON^D \rangle, \\ for\ i = 1,2,.....m \text{ and } \ j = 1,2,.....n. \end{Bmatrix}$$

Convert the W^C in to a new matrix as a single number representation using the below equation.

$$K_{ij}^c = \begin{Bmatrix} card(k) / c \in C \wedge k \in W_{ij}^c, \\ for\ i = 1,2,.....m \text{ and } j = 1,2,.....n \end{Bmatrix}$$

Total number of items with equal attribute values in ON^c

$$CI_i^c = \sum_{j=1}^{n} K_{ij}^c \text{ for } i = 1,2,.....m$$

Total number of items with equal attributes values in ON^D

$$DI_j^c = \sum_{i=1}^{m} K_{ij}^c \text{ for } for\ j = 1,2,.....n$$

The numerical count of all items in the set is

$$TI^c = \sum_{i=1}^{m} \sum_{j=1}^{n} K_{ij}^c$$

The confidence factor for each conditional and decision value pair is

$$Confidence_{ij}^C = \frac{K_{ij}^c}{CI_i^c} \text{ for } i = 1,2,.....m \text{ and } j = 1,2,.....n$$

The confidence factor for each decision value independent of any conditional value is

$$Confidence_j^D = \frac{DI_j^c}{TI^c} \text{ for } j = 1,2,\ldots\ldots\ldots\ldots n.$$

The Coverage factor value for each indiscernible conditional value and decision value pair is

$$Coverage_{ij}^c = \frac{K_{ij}^c}{DI_j^c} \text{ for } i = 1,2\ldots\ldots\ldots\ldots m;\ j = 1,2\ldots\ldots\ldots\ldots n.$$

The difference between the confidence dependency matrix and confidence independent matrix of any conditional value is defined as the gain factor (G).

$$Gain = G_{ij}^c = Confidence_{ij}^C - Confidence_j^D$$

For:i=1,2,…….m; j=1,2,…….n

The positive value of gain factor indicates the increase in the upper approximation. The negative value indicates the decrease in the upper approximation.

The gain factor alone is not enough to describe the gain factor can be scaled up suitably to reflect the hold up for each decision class, hence rough information measure (RIM) is defined, which is the product of gain factor and coverage factor.

$$RIM_{ij}^c = G_{ij}^c \times Coverage_{ij}^c;\ for\ i = 1,2,\ldots\ldots\ldots\ldots m;$$
$$j = 1,2,\ldots\ldots\ldots\ldots n$$

From this rough information measure matrix the breaking points are identified. The breaking point is a point at which the rough information measure changes from one attribute to another attribute when one attribute values are arranged in sorted order in the rough information measure matrix. At each breaking point the rough information sum is calculated. The attribute with maximum rough information measure (RIM) sum is selected as decision attribute and the classification rules are generated using the corresponding lower and upper attribute values.

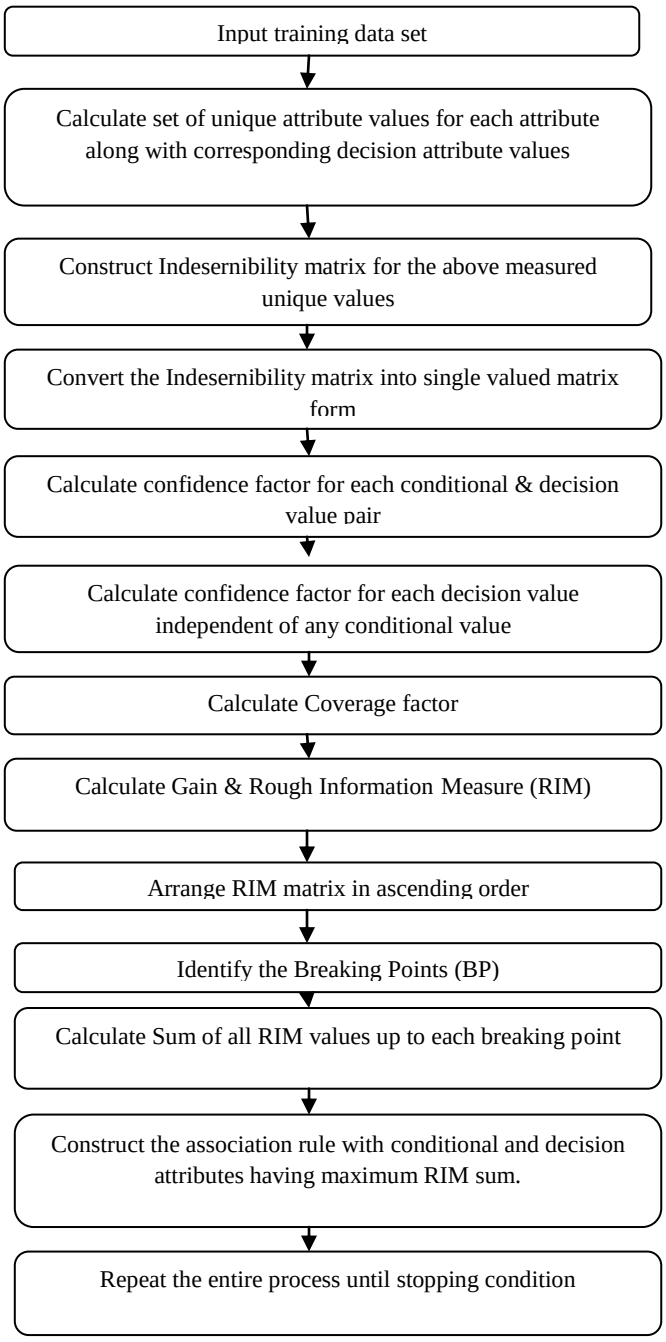

Figure 1 .Model flow chart

4. Stopping criterion

- If all the items in the data set belong to the same class
- If all items have similar records in the attribute list
- If the dataset is empty

5. Performance metrics

The widely used performance metrics in classification techniques are accuracy, sensitivity and specificity. The ability of the method to correctly detect the class label of previously unseen or new data is defined as the accuracy of the model. Sensitivity measures the ability to identify the occurrence of target class accurately. The ability of the method to separate the target class is defined as the specificity of the method (14). The accuracy, sensitivity and specificity measures are shown in table 3.

Table 3.Performance Measures

Measure	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$

Where

True positives (TP): True positives refers to the positive tuples that were correctly labeled by the classifier.

True negatives (TN): True negatives refers to the negative tuples that were correctly labeled by the classifier.

False positives (FP): False positives to the negative tuples that were incorrectly labeled as positive.

False negatives (FN): False negatives refers to the positive tuples that were mislabeled as negative.

6. k-fold cross validation

k-fold cross validation is the adequate quantity for classifier performance. k-fold cross-validation technique segments the whole data into k fold where each fold will be taken into consideration for both training and testing the method (15). This process is repeated k times so that each segment is used for testing exactly once. The average accuracy of the algorithm is obtained by taking average of the k different test results. In this work we used three fold cross validation to test the performance of the model.

7. Results

In this paper the authors proposed a computational intelligence technique using rough set theory for the diagnosis of diabetes disease. The method performance was verified using three fold cross validation technique in each fold 224 training items and 112 test items are used for training and testing the algorithm. MATLAB software is used to implement the method. The performance of the technique is evaluated using the confusion matrix obtained in three training samples as shown in below figures 2, 3 and 4. The accuracy, sensitivity and specificity of the method for the three fold are presented in table 4.

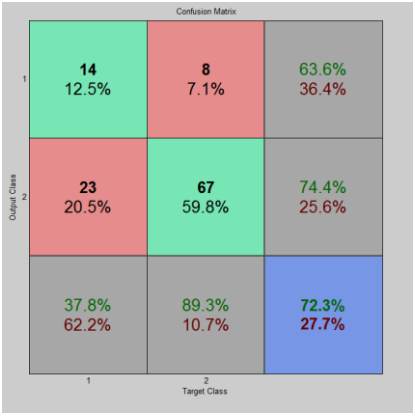

Figure 2. Confusion matrix for first test set

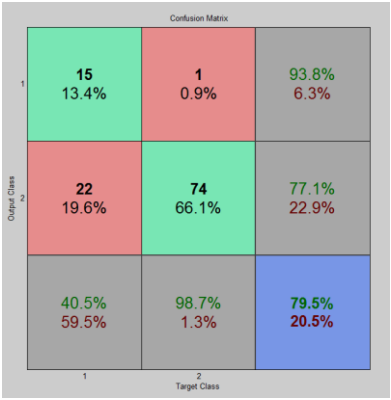

Figure 3. Confusion matrix for second test set

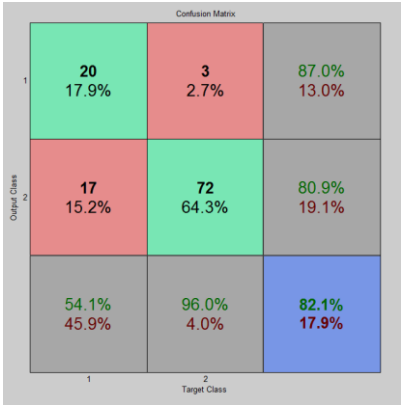

Figure 4. Confusion matrix for second test set

Table 4. Summary of the confusion matrix for both training and testing

Fold	Testing		
	Accuracy	**Sensitivity**	**Specificity**
1	72.3	37.8	89.3
2	79.5	40.5	98.7
3	82.1	96.0	54.1

Table 5. Classification Accuracy

Method	Author	Accuracy (%)	Ref.
SLIQ	Mehta.M, R. Agarwal, and J. Rissanen (1996)	67.92	(16)
RBF	Kayaer.Yildirim (2003)	68.23	(17)
CART	Ster. Dbobnikar (1996)	72.8	(18)
MLP	Ster. Dbobnikar (1996)	75.2	(18)
Naïve	Bayes Friedman (1997)	74.5	(19)
C 4.5	J.R. Quinlan (1993)	65.06	(20)
GGFSDT	K V S R P Varma et al. (2013)	75.8	(21)
RST	Present Method (2014)	**77.9**	

8. Results

In this work a computational intelligence technique was proposed for the early and effective diabetes detection using rough set theory using indiscernible matrix and rough information measure. The work was carried out with 336 PID data samples taken from UCIMLR. The average test accuracy obtained with this model is 77.9% on PID data set. In future work planning to extend this work with other disease data sets.

9. Acknowledgement

Allam Appa Rao acknowledges the financial support of DST-IRHPA Scheme vide Lr No. IR/SO/LU/03/2008/1 dated 24-12-2010.

10. References

[1] From data mining to knowledge descovery: an overview. Fayyad, U, Piatetsky, Shapiro G and Smyth, P. 1996a, Advances in knowledge discovery and data mining.

[2] The INterestingness of deviations. Piatetsky-Shapiro, G and Matheus, C J. Montreal-Canada : AAAIPressm Menlo Park-USA, 1995. International Conference on Knowledge Discovery and Data Mining. pp. 23-36.

[3] Fuzzy sets Information and Control. Zadeh, L A. 1965, Vol. 8, pp. 338-353.

[4] Rough Sets. Pawlak, Z. 1982, International Jourrnal of computer Information Sciences, Vol. 11, pp. 341-356.

[5] Dariusz, G Mikulski. Rough set based splitting criterion for binary decision tree classifiers. 2006.

[6] Rough Sets Perspective on Data and Knowledge. Komorowski, J, et al. New York-USA : Oxford University Press, 1999, The Handbook of Data Mining and Knowledge Discovery, pp. 134-149.

[7] Rough Sets in Hybrid Methods for Pattern Recognition. Mrozek, A and Cyran, K. 2, February 2001, International Journal of INtelligence Systems, Vol. 16, pp. 149-168.

[8] Assessment of Conert Hall Acoustics using Rough Set and Fuzzy Set Approach. Kostek, B. [ed.] S Pal and A Skowron. Secaucus-USA : Springer-Verlag Co., 1999, Rough Fuzzy Hybridization: A New Trend in Decision-Making, pp. 318-396.

[9] Power System Security Analysis based on Rough Classificaion. Lambert-Torres, G, et al. [ed.] S Pal and A Skowron. Secaucus-USA : Springer-Verloag Co., 1999, Rough Fuzzy Hybridizaiton: A New Trend in Decision-Making, pp. 263-300.

[10] Rough Set Approach to Sunspot Classification Problem. Nguyen, S.H, Nguyen, T.T and Nguyen, H.S. Regina-Canada : Springer, Secaucus-USA, 2005. International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing- Lecture Notes in Artificail Intelligence 3642. pp. 263-272.

[11] RST-Based System Design of Hybrid Intelligent Control . Xie, G, Wang, F and Xie, K. The Hague- The Netherlands : IEEE Press, New Jersey-USA, 2004. IEEE International Confernece on Systems, Man and Cybernetics. pp. 5800-5805.

[12] Integration Method of Ant Colony Algorithm and Rough Set Theory for Simultaneous Real Value Attribute Discretization and Attribute Reduction. He, Y, Chen, D and Zhao, W. [ed.] F T S Chan and M K Tiwari. Budapest-Hungary : I Tech Education and Publishing, 2007, Swarm Intelligence: Focus on Ant and Particle Swarm Optimization, pp. 15-36.

[13] Diabetes incidence and prevalence in Pima Indians: A 19-fold greater incidence than in Rochester, Minnesota. William, C Knowler. 6 1978, American Journal of Epidemology, Vol. 108, pp. 497-505.

[14] Han, Jiawei, Kamber, Micheline and Pei, Jian. Data Mining Concepts and Techniques. Waltham : MORGAN KAUFMANN, 2012.

[15] Predicting breast cancer data survivabiltiy: A comparision of three dataming methods. Delen, D,G Walker and Kadam, A. 2 2005, Artificial Intelligence in Medicine, Vol. 34, pp. 113-127.

[16] SLIQ: A fast scalable classifier for data mining in Extending Database Technology. Mehta, M, Agrawal, R and Riassnen, J. Avignon, France : s.n., 1996, Springer, pp. 18-32.

[17] Medical diagnosis on Pima Indian diabetes using general regression neural networks. Kayaer, K and Yildirim, T. Istanbul : s.n., 2003. International Conference on artificial neural networks and neural information processing. pp. 181-184.

[18] Neural networks in medical diagnosis: comparison with other methods. Ster, B and Dobnikar, A. London : s.n., 1996. International Conference on Engineering Applications with Neural Networks. pp. 427-430.

[19] Bayesian networks classifiers. Friedman, N, Geiger, D and Goldszmit, M. 1997, Machine Learning, Vol. 29, pp. 131-163.

[20] Quinlan, J R. Programs for Machine Learning. San Mateo, California : Morgan Kaufmann, 1993.

[21] A computational intelligence approach for a better diagnosis of diabetic patients. Kamadi, Varma V S R P, et al. August 2013, Computers and Electrical Engineering.