

# **Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?**

**Joseph L. Breault, MD, MPH, MS**

Department of Health Systems Management, Tulane University  
Department of Family Practice, Alton Ochsner Medical Foundation  
joebreault@tulanealumni.net

**ABSTRACT:** The publicly available Pima Indian diabetic database (PIDD) at the UC-Irvine Machine Learning Lab has become a standard for testing data mining algorithms to see their accuracy in predicting diabetic status from the 8 variables given. Looking at the 392 complete cases, guessing all are non-diabetic gives an accuracy of 65.1%. Since 1988, many dozens of publications using various algorithms have resulted in accuracy rates of 66% to 81%. Rough sets as a data mining predictive tool has been used in medical areas since the late 1980s, but not applied to the PIDD to our knowledge. When we apply rough sets to PIDD using ROSETTA software, there are many different options within the software to choose from. The predictive accuracy was 73.8% with a 95% CI of (71.3%, 76.3%) with one of the methods we used. Rough sets are a useful addition to the analysis of diabetic databases.

**Keywords:** Diabetes, Pima Indians, Data Mining, Diabetes, Rough Sets, Knowledge Discovery in Databases, ROSETTA

## **INTRODUCTION TO DATA MINING DIABETIC DATABASES**

There has been extensive work on diabetic registries for a variety of purposes. Databases have been used to query for diabetes (Michel and Beguin 1994), as a comprehensive management tool to improve diabetic care and communications among professionals (Flack 1995; Kelling, Wentworth et al. 1997), and to provide continuous quality improvement in diabetes care (Kopelman and Sanderson 1996). The Veterans Administration (VA) developed their diabetic registry from an outpatient pharmacy database and matched social security numbers to add VA hospital admission data to it. They identified 139,646 veterans with diabetes (Pogach, Hawley et al. 1998). The Belgian Diabetes Registry was created by required reporting of all incident cases of type 1 diabetes and their first degree relatives younger than 40. This has facilitated epidemiologic and genetic studies (Dorchy 1999). One British hospital linked their 7000 patient database to their National Health Services Central Registry to identify mortality data and found that diabetes was recorded in only 36% of death certificates, so analysis of death certificates alone gives poor information about mortality in diabetes (Weng, Coppini et al. 1997).

Diabetes is a particularly opportune disease for data mining technology for a number of reasons. First, because the mountain of data is there. Second, diabetes is a common disease that costs a great deal of money, and so has attracted managers and payers in the never ending quest for saving money and cost efficiency. Third, diabetes is a disease that can produce terrible complications of blindness, kidney failure, amputation, and premature cardiovascular death, so physicians and regulators would like to know how to

improve outcomes as much as possible. Data mining might prove an ideal match in these circumstances.

## **THE PIMA INDIAN DIABETIC DATABASE**

The Pima Indians may be genetically predisposed to diabetes (Hanson, Ehm et al. 1998), and it was noted that their diabetic rate was 19 times that of a typical town in Minnesota (Knowler, Bennett et al. 1978). The National Institute of Diabetes and Digestive and Kidney Diseases of the NIH originally owned the Pima Indian Diabetes Database (PIDD). In 1990 it was received by the UC-Irvine Machine Learning Repository and can be downloaded at [www.ics.uci.edu/~mlearn/MLRepository.html](http://www.ics.uci.edu/~mlearn/MLRepository.html). The database has  $n=768$  patients each with 9 numeric variables: (1) number of times pregnant, (2) 2-hour OGTT plasma glucose, (3) diastolic blood pressure, (4) triceps skin fold thickness, (5) 2-hour serum insulin, (6) BMI, (7) diabetes pedigree function, (8) age, (9) diabetes onset within 5 years (0, 1). The goal is to use the first 8 variables to predict #9. There are 500 non-diabetic patients (class = 0) and 268 diabetic ones (class = 1) for an incidence rate of 34.9%. Thus if you simply guess that all are non-diabetic, your accuracy rate is 65.1% (or error rate of 34.9%). We expect a useful data mining or prediction tool to do much better than this.

There are a few errors in the data. Although the database is labeled as having no missing values, someone liberally added zeros where there were missing values. Five patients had a glucose of 0, 11 more had a body mass index of 0, 28 others had a diastolic blood pressure of 0, 192 others had skinfold thickness readings of 0, and 140 others had serum insulin levels of 0. None of these are physically possible, and after they were deleted there were 392 cases with no missing values. Studies that did not realize the previous zeros were in fact missing variables essentially used a rule of substituting zero for the missing variables. Ages range from 21 to 81 and all are female.

## **STUDIES ON THE PIMA INDIAN DIABETIC DATABASE**

There have been many studies applying data mining techniques to the PIDD. The independent or target variable is diabetes status within 5 years, represented by the 9<sup>th</sup> variable (class=1).

Smith et al. used a neural network ADAP algorithm using Hebbian learning to build associative models. They used 576 randomly selected cases for training and the remaining 192 test cases showed an accuracy of 76% (Smith, Everhart et al. 1988).

Quinlan applied C4.5 and it was 71.1% accurate (Quinlan 1993).

Wahba's group at the University of Wisconsin applied penalized log likelihood smoothing spline analysis of variance (PSA). They eliminated patients with glucose and BMIs of zero leaving  $n=752$ . They used 500 for the training set, and the remaining 252 as the evaluation set which showed an accuracy of 72% for the PSA model and 74% for a GLIM model (Wahba, Gu et al. 1992).

Michie et al. used 22 algorithms with 12-fold cross validation and reported the following accuracy rates on the test set: Discrim 77.5%, Quaddisc 73.8%, Logdisc 77.7%, SMART 76.8%, ALLOC80 69.9%, k-NN 67.6%, CASTLE 74.2%, CART 74.5%, IndCART 72.9%, NewID 71.1%, AC<sup>2</sup> 72.4%, Baytree 72.9%, NaiveBay 73.8%, CN2 71.1%, C4.5 73%, Itrule 75.5%, Cal5 75%, Kohonen 72.7%, DIPOL92 77.6%, Backprop 75.2%, RBF 75.7%, and LVQ 72.8% (Michie, Spiegelhalter et al. 1994, p. 158).

The Multi-Stream Dependency Detection (MSDD) algorithm was used on two-thirds of the dataset for training. No mention is made of deleting any missing values. Accuracy on the one-third for evaluation was 71.33% (Oates 1994).

Turney applied algorithms (ICET, EG2, CS-ID3, IDX) to the dataset with cost information to determine which algorithm is best when including the costs of classification, tests and classification errors. ICET performed the best (Turney 1995).

Bayesian neural nets were applied to the dataset using the same deletions and training sample as Wahba. The standard neural network had an accuracy of 75.4%, the Bayesian approach 79.5% (Bioch, van der Meer et al. 1996).

Ripley used the 532 cases that excluded missing insulin levels, with a training set of 200 and a test set of 332. When methods could deal with missing values, he added 100 of the missing insulin cases to the training set. Accuracy rate for logistic regression was 80.2%, MARS and PPR models 77.4%, neural network 77%, k-NN for k=9 75.3%, OLVQ 78.9%, CART 75.6% increasing to 77.7% if 100 incomplete cases were added to the training set (Ripley 1996).

The ARTMAP-IC neural network adds distributed prediction and category instance counting to the basic fuzzy ARTMAP system. This was applied to the database with the same training and test sets as Smith. It achieved an accuracy of 81%, when ARTMAP was 66%, logistic regression 77%, and KNN 77% (Carpenter and Markuzon 1998).

Khan used multiplier-free feedforward networks (MFN), and correctly noted that 49% of the patients had zero values for variables that cannot be zero. Nevertheless, he selected all of the diabetic patients and an equal number of non-diabetics to attain a balanced set but with missing variables. It is not clear whether the non-diabetics were randomly selected or selected to minimize missing variables. He took half of these in the form of a balanced subgroup (n=268) as a training set and standardized the 8 variables to zero mean and unit variance. He computed accuracy for the MFN and also discrete-weight networks (DWN) and continuous-weight networks (CWN) using the n=268 evaluation set and results were 78.0%, 76.9%, and 78.4% respectively (Khan 1998).

Eklund & Hoang used a number of algorithms on the dataset with 80% training/20% evaluation sets. The problem of missing variables set to zero was ignored. The accuracy of the algorithms tested was C4.5 71.02%, C4.5 rule 71.55%, ITI 73.16%, LMDT 73.51%, and CN2 72.19% (Eklund and Hoang 1998).

Liu integrated classification and association rule mining in class association rules (CAR) that were applied to the data set. The best of the 4 CAR models had an accuracy of 73.1% compared to 75.5% for C4.5 rules. In many other datasets it tended to outperform C4.5 (Liu 1998).

King et al. used 14 algorithms on the PIDD. They discarded the insulin variable with the most missing cases, leaving n=532. The accuracy of the data mining tools used was CART 76%, Scenario 30%, See5 73%, S-Plus 79%, WizWhy 74%, DataMind 69%, DMSK 67%, NeuroShell2--Neural 77%, PcOLPARS 81%, PRW 80%, MQ Expert 77%, NeuroShell2—PolyNet 78%, Gnosis 81%, and KnowledgeMiner 78% (King, Elder IV et al. 1998).

Classification by aggregating emerging patterns (CAEP) applied to the PIDD was initially 72% accurate, but could only identify 30% of the diabetic patients. After modifications it was 75% accurate (Wong 2000).

Although the cited articles use somewhat different subgroups of the PIDD, accuracy for predicting diabetic status ranges from 66% to 81%. While some of these are means of a larger group of randomizations, most are simply one randomization into a training set and test set arriving at an accuracy. This run the risk of a particularly good or bad accuracy being a quirk of that particular randomization rather than the method used.

## **ROUGH SETS IN MEDICAL DATA ANALYSIS**

Rough sets investigate structural relationships in the data rather than probability distributions, and produce decision tables rather than trees (Ziarko 1991). This method forms equivalence classes within the training data, approximating it with a class below it

and a class above it. A variety of algorithms can be used to define the classification boundaries. Rough sets also do feature reduction. Finding minimal subsets (reducts) of attributes that are efficient for rule making is a central part of its process (Han and Kamber 2001, p. 316).

Rough sets have been applied to peritoneal lavage in pancreatitis (Slowinski, Slowinski et al. 1988), toxicity predictions (Hashemi, Jelovsek et al. 1993), development of medical expert system rules (Tsumoto and Tanaka 1994; Tsumoto and Tanaka 1995), prediction of death in pneumonia (Paterson 1995), identification of patients with chest pain who do not need expensive additional cardiac testing (Øhrn, Vinterbo et al. 1997; Komorowski and Øhrn 1999), diagnosing congenital malformations (Tsumoto 1998), prediction of relapse in childhood leukemia (Podraza and Podraza 1999), and to predict ambulation in people with spinal cord injury (Øhrn and Rowland 2000). There are extensive reviews of their use in medicine (Komorowski, Polkowski et al. 1998; Øhrn 1999). To our knowledge, there are no publications about their application to the PIDD.

A recent study used a dataset of 107 children with diabetes from a Polish medical school. Attributes available for each child were sex, age in years at diagnosis (<7, 7-12, 13-15, >15), disease duration in years (<6, 6-10, >10), family history (yes/no), respiratory infection (yes/no), Readmission (yes/no), type of insulin used, HgbA1c (<8, 8-10, >10), and microalbuminuria (yes/no). Rough set techniques were applied and decision rules generated to predict microalbuminuria. The best predictor was age < 7 predicting no microalbuminuria 83.3% of the times, followed by age 7-12 with disease duration 6-10 predicting microalbuminuria 80.8% of the times (Stepaniuk 1998). In a follow-up paper using the same dataset, the author lists additional attributes including hypertension (yes/no), body mass (<3, 3-97, >97), high cholesterol (yes/no), and high triglyceride (yes/no). The importance of the attributes to prediction of microalbuminuria was investigated by 3 methods: reducts, significance, and the wrapper approach. Rough set methods showed the highest accuracy (77%) using 9 attributes, while the wrapper method got best results (79.4%) with 6 attributes (Stepaniuk 1999).

Those interested in a recent review of rough set theory as a method for data mining can see chapter 2 of (Cios, Pedrycz et al. 1998)

## **ROUGH SETS APPLIED TO THE PIMA INDIAN DIABETIC DATABASE**

We randomly divided the 392 complete cases in the PIDD into a training set (n=300), and a test set (n=92). The data is available in an Microsoft Excel spreadsheet elsewhere in this volume, as test0 and training0. The ROSETTA software<sup>1</sup> was downloaded from [www.idi.ntnu.no/~aleks/rosetta/](http://www.idi.ntnu.no/~aleks/rosetta/) (Øhrn and Komorowski 1997). We used ROSETTA GUI ver 1.4.40, kernel version 1.0, RSES ver 1.41, release build 22:41:48 Nov 8 2000.

The following steps use the 300 case training set.

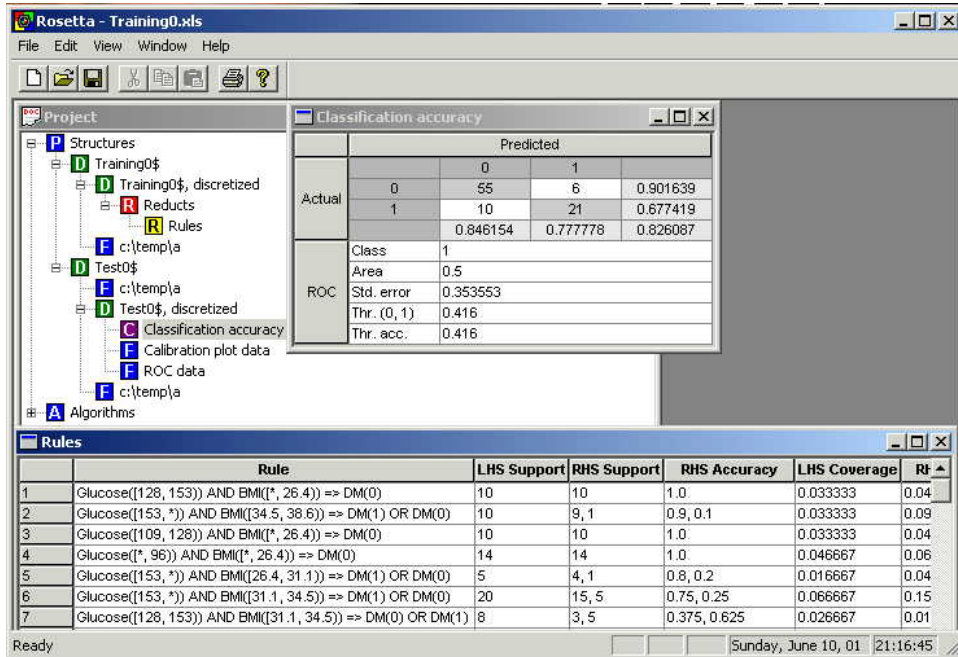
ROSETTA's first step after imputing data is to deal with missing values in one of 5 ways, but we had already removed these.

Next is the discretization step, where each variable is divided into a limited number of value groups (e.g., age groups instead of the specific age). There are 9 ways to do this and we chose the equal frequency binning criteria with k=5 bins.

---

<sup>1</sup> The homepage states "ROSETTA is a toolkit for analyzing tabular data within the framework of rough set theory, and runs on PCs operating under Windows NT/98/95/2000. Computational kernel and GUI front-end designed and implemented at the Knowledge Systems Group, Dept. of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway. Sections of the computational kernel (RSES) developed at the Group of Logic, Inst. of Mathematics, University of Warsaw, Poland."

The next step is creating reducts, which are subset vectors of attributes that facilitate rule generation with minimal subsets. This can be done by 8 methods; we choose the Johnson reducer algorithm. Options selected for this algorithm were discernibility = object related (universe = all objects), table interpretation = Modulo with boundary thinning 0.1, no checks in the discernibility predicate or memory usage options, and advanced parameters using approximate solutions with a hitting fraction of 0.95. The software generated rules can then be used on the test set.



**Figure 1** Image of ROSETTA software on training0 and test0 data sets

The next step is applying a classification method to the test set and we choose the batch classifier with the standard/tuned voting method (RSES). Options clicked are a fallback to 0 (non-diabetic) with a probability of 0.651, selecting the best according to the classifier's measure of certainty, and under parameters using the rule generated with the training set, majority voting, and excluding rules with normalized support below 0. When the generated training rules are applied to the test set of 92 cases the prediction accuracy is 82.6%, which is better than all of the previous machine learning algorithms noted above. The "confusion matrix" produced by the program is below (1=diabetes, 0=non diabetes).

When the discretization step was tweaked by domain knowledge (selecting 5 intervals for each variable based on being most clinically meaningful), results looked slightly improved on the training set (91.7% vs. 91.0%), but were much worse on the test set (75.0% vs. 82.6%).

In exploring which discretization method works best on the training set for the Johnson algorithm with tuned voting, we had the following accuracies: Boolean 96%, entropy 78%, binning (k=5) 91%, naïve 100%, semi-naïve 99%, and BooleanRSES 90%. We suspected that the ones in the high 90s are overfitted and would not do as well on the test set, thus binning might be a good choice. Test results were Boolean 66%, entropy 62%, binning (k=5) 83%, naïve 67%, semi-naïve 78%, and BooleanRSES 74%.

A final issue is the binning number that works best. On the training set if we use k=2, 3, 4, 5, 6 and 7 we get the following accuracies using the Johnson reduct with tuned voting: 81.3%, 90.3%, 87.3%, 91.0%, 91.3%, and 95%. We suspect the highest binning numbers are heading toward overfitting. When the various binning numbers are used on the test set, we get accuracies of 76.1%, 79.3%, 81.5%, 82.6%, 78.3%, 81.5% indicating k=5 works best.

### USING A SERIES OF RANDOM SAMPLES TO GET A MEAN AND CI

The 82.6% accuracy rate is surprisingly good, and exceeds the previously used machine learning algorithms that ranged from 66-81%. To make sure this is not just a quirk of the particular random sample that we obtained, 9 additional random samples were obtained from the 392 cases. All were divided into a training set of 300 and a test set of 92, and these are in the Microsoft Excel file elsewhere in this volume. Using the same method as above, the confusion matrices are listed in Figure 2.

		Predicted					Predicted		
			0	1			0	1	
Actual	0	55	6	0.901639	Actual	0	51	16	0.761194
	1	10	21	0.677419		1	8	17	0.68
		0.846154	0.777778	0.826087			0.864407	0.515152	0.73913
sample 0 from previous section									
		Predicted					Predicted		
			0	1			0	1	
Actual	0	50	16	0.757576	Actual	0	45	12	0.789474
	1	10	16	0.615385		1	18	17	0.485714
		0.833333	0.5	0.717391			0.714286	0.586207	0.673913
		Predicted					Predicted		
			0	1			0	1	
Actual	0	55	16	0.774648	Actual	0	51	7	0.87931
	1	4	17	0.809524		1	13	21	0.617647
		0.932203	0.515152	0.782609			0.796875	0.75	0.782609
		Predicted					Predicted		
			0	1			0	1	
Actual	0	46	12	0.793103	Actual	0	48	11	0.813559
	1	17	17	0.5		1	14	19	0.575758
		0.730159	0.586207	0.684783			0.774194	0.633333	0.728261
		Predicted					Predicted		
			0	1			0	1	
Actual	0	44	22	0.666667	Actual	0	55	11	0.833333
	1	11	15	0.576923		1	13	13	0.5
		0.8	0.405405	0.641304			0.808824	0.541667	0.73913

**Figure 2 Confusion matrices from ROSETTA for the 10 samples**

From the 10 randomizations, the mean is 73.2% with a 95% confidence interval (CI) of (69.2% - 77.2%) as detailed in the SPSS printout in Figure 3. It thus turns out that our initial random sample (training0 and test0 data sets) gave an unusually good predictive accuracy of 82.6% by chance and this is beyond the 95% CI. This is also a reminder that in evaluating data mining programs it is important to include multiple random samples (or 10-fold cross validation, etc.) to insure that results are not just by chance.

### OTHER ALTERNATIVES IN ROSETTA

The software has many options and choices. For example, using binning with k=5, reducing with the exhaustive calculation (RSES) option with objected related all objects, and Modulo. We generate the rules on the 10 training sets. Then with the respective test

sets, we classify them using the standard/tuned voting (RSES) with its defaults. The 10 accuracies ranged from 68.5% to 79.3% with a mean of 73.9% and a 95% CI of (71.5%, 76.3%).

### Descriptives

			Statistic	Std. Error
ACCURACY	Mean		.73152170	1.77E-02
	95% Confidence Interval for Mean	Lower Bound	.69152777	
		Upper Bound	.77151563	
	5% Trimmed Mean		.73128017	
	Median		.73369550	
	Variance		3.126E-03	
	Std. Deviation		5.59E-02	
	Minimum		.641304	
	Maximum		.826087	
	Range		.184783	
	Interquartile Range		.10054350	
	Skewness		.066	.687
	Kurtosis		-.415	1.334

**Figure 3 SPSS printout of mean and descriptives for accuracy of 10 samples**

A slightly newer version was recently released (GUI ver 1.4.41, kernel ver 1.01, release build 22:13:30 May 27 2001). Using this version, the binning with  $k=5$ , and the defaults for the Johnson reducer algorithm, rules were constructed for each of the 10 randomizations of the PIDD training sets from above. This time the test sets were classified according to defaults of the naïve Bayes classifier, and the 10 accuracies ranged from 69.6% to 85.5% with a mean of 73.8% and a 95% CI of (71.3%, 76.3%).

Although the software is easy to use, it will take time to explore the various options to see if these numbers can be significantly improved on.

## CONCLUSIONS

Rough sets and the Rosetta software are useful additions to the analysis of diabetic databases. The accuracy of predicting diabetic status on the PIDD was 82.6% on the initial random sample, which exceeds the previously used machine learning algorithms that ranged from 66-81%. Using a group of 10 random samples the mean accuracy was 73.2% with a 95% CI of (69.2% - 77.2%). Other methods within the software gave a mean accuracy of 73.9% and a 95% CI of (71.5%, 76.3%), and a mean of 73.8% with a 95% CI of (71.3%, 76.3%). The ROSETTA software is easy to use and get good results from quickly, but it may take some experience before the proper set of options to get optimal results is discovered.

## REFERENCES

- Bioch, J. C., O. van der Meer, et al. (1996). Classification using Bayesian neural nets. The 1996 IEEE International Conference on Neural Networks, p. 1488-1493, Washington, DC, Institute of Electrical and Electronics Engineers.

- Carpenter, G. A. and N. Markuzon (1998). "ARTMAP-IC and medical diagnosis: instance counting and inconsistent cases." Neural Networks **11**(2): 323-336.
- Cios, K. J., W. Pedrycz, et al. (1998). Data mining methods for knowledge discovery. Boston, Kluwer Academic.
- Dorchy, H. (1999). "[Screening, prediction and prevention of type 1 diabetes. Role of the Belgian Diabetes Registry]." Rev Med Brux **20**(1): 15-20.
- Eklund, P. W. and A. Hoang (1998). Classifier Selection and Training Set Features: LMDT, [citeseer.nj.nec.com/309003.html](http://citeseer.nj.nec.com/309003.html).
- Flack, J. R. (1995). "Seven years experience with a computerized diabetes clinic database." Medinfo **8**(Pt 1): 332.
- Han, J. and M. Kamber (2001). Data mining: concepts and techniques. San Francisco, Morgan Kaufmann Publishers.
- Hanson, R. L., M. G. Ehm, et al. (1998). "An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians." Am J Hum Genet **63**(4): 1130-1138.
- Hashemi, R. R., F. R. Jelovsek, et al. (1993). "Developmental toxicity risk assessment: a rough sets approach." Methods Inf Med **32**(1): 47-54.
- Kelling, D. G., J. A. Wentworth, et al. (1997). "Diabetes mellitus. Using a database to implement a systematic management program." N C Med J **58**(5): 368-371.
- Khan, A. H. (1998). Multiplier-free Feedforward Networks, [citeseer.nj.nec.com/6034.html](http://citeseer.nj.nec.com/6034.html). **1998**.
- King, M. A., J. F. Elder IV, et al. (1998). Evaluation of Fourteen Desktop Data Mining Tools. IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, [citeseer.nj.nec.com/293388.html](http://citeseer.nj.nec.com/293388.html).
- Knowler, W. C., P. H. Bennett, et al. (1978). "Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota." Am J Epidemiol **108**(6): 497-505.
- Komorowski, J. and A. Øhrn (1999). "Modeling prognostic power of cardiac tests using rough sets." Artif Intell Med **15**(2): 167-191.
- Komorowski, J., L. Polkowski, et al. (1998). Rough sets: a tutorial. Rough-fuzzy hybridization: a new trend in decision-making. S. K. Pal and A. Skowron. New York, Springer Verlag: xiv, 454.
- Kopelman, P. G. and A. J. Sanderson (1996). "Application of database systems in diabetes care." Med Inform (Lond) **21**(4): 259-271.
- Liu, B. (1998). Integrating Classification and Association Rule Mining. KDD-98, Knowledge Discovery and Data Mining, New York: 80-86.
- Michel, C. and C. Beguin (1994). "Using a database to query for diabetes mellitus." Stud Health Technol Inform **14**: 178-182.
- Michie, D., D. J. Spiegelhalter, et al. (1994). Machine learning, neural and statistical classification. New York, Ellis Horwood.
- Oates, T. (1994). MSDD as a Tool for Classification. EKSL Memorandum 94-29, Department of Computer Science, University of Massachusetts at Amherst.

- Øhrn, A. (1999). Discernibility and Rough Sets in Medicine: Tools and Applications. Department of Computer and Information Science. Trondheim, Norway, Norwegian University of Science and Technology: 239.
- Øhrn, A. and J. Komorowski (1997). ROSETTA: A Rough Set Toolkit for Analysis of Data. Joint Conference of Information Sciences: semiotics, fuzzy logic, soft computing, computer vision, neural computing, genetic algorithm, pattern recognition, evolutionary computing, Durham, NC, Duke University Press: 403-407.
- Øhrn, A. and T. Rowland (2000). "Rough sets: a knowledge discovery technique for multifactorial medical outcomes." Am J Phys Med Rehabil **79**(1): 100-108.
- Øhrn, A., S. Vinterbo, et al. (1997). "Modelling cardiac patient set residuals using rough sets." Proc AMIA Annu Fall Symp: 203-207.
- Paterson, G. I. (1995). "A rough sets approach to patient classification in medical records." Medinfo **8**(Pt 2): 910.
- Podraza, W. and H. Podraza (1999). "Childhood leukaemia relapse risk factors. A rough sets approach." Med Inform Internet Med **24**(2): 91-108.
- Pogach, L. M., G. Hawley, et al. (1998). "Diabetes prevalence and hospital and pharmacy use in the Veterans Health Administration (1994). Use of an ambulatory care pharmacy-derived database." Diabetes Care **21**(3): 368-373.
- Quinlan, J. R. (1993). C4.5: programs for machine learning. San Mateo, Calif., Morgan Kaufmann Publishers.
- Ripley, B. D. (1996). Pattern recognition and neural networks. Cambridge ; New York, Cambridge University Press.
- Slowinski, K., R. Slowinski, et al. (1988). "Rough sets approach to analysis of data from peritoneal lavage in acute pancreatitis." Med Inform (Lond) **13**(3): 143-159.
- Smith, J. W., J. E. Everhart, et al. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Symposium on Computer Applications and Medical Care (Washington, DC). R. A. Greenes. Los Angeles, CA, IEEE Computer Society Press: 261-265.
- Stepaniuk, J. (1998). Rough Set Based Data Mining in Diabetes Mellitus Data Table. EUFIT '98, September 7-10, 1998: Intelligent techniques and soft computing, Aachen, Germany, Verlag Mainz: 980-984.
- Stepaniuk, J. (1999). Rough set data mining of diabetes data. Foundations of intelligent systems: 11th International Symposium, ISMIS'99, Warsaw, Poland, June 8-11, 1999: proceedings. Z. Ras and A. Skowron. Berlin; New York, Springer: 457-465.
- Tsumoto, S. (1998). "Automated knowledge acquisition from clinical databases based on rough sets and attribute-oriented generalization." Proc AMIA Symp: 548-552.
- Tsumoto, S. and H. Tanaka (1994). "Induction of medical expert system rules based on rough sets and resampling methods." Proc Annu Symp Comput Appl Med Care: 1066-1070.
- Tsumoto, S. and H. Tanaka (1995). "Induction of expert system rules based on rough sets and resampling methods." Medinfo **8**(Pt 1): 861-865.

- Turney, P. D. (1995). "Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm." Journal of Artificial Intelligence Research **2**: 369-409.
- Wahba, G., C. Gu, et al. (1992). Soft Classification, a.k.a. Risk Estimation, via Penalized Log Likelihood and Smoothing Spline Analysis of Variance. The mathematics of generalization: the proceedings of the SFI/CNLS Workshop on Formal Approaches to Supervised Learning, Santa Fe, Addison-Wesley Pub. Co.: 331-360.
- Weng, C., D. V. Coppini, et al. (1997). "Linking a hospital diabetes database and the National Health Service Central Register: a way to establish accurate mortality and movement data." Diabet Med **14**(10): 877-883.
- Wong, L. (2000). Datamining: Discovering Information from Bio-Data, [citeseer.nj.nec.com/375806.html](http://citeseer.nj.nec.com/375806.html).
- Ziarko, W. (1991). The discovery, analysis, and representation of data dependencies in databases. Knowledge discovery in databases. G. Piatetsky-Shapiro and W. Frawley. Menlo Park, Calif., AAAI Press: MIT Press: 195-209.