A Comparison of Different Decision Algorithms Used in Volumetric Storm Cells Classification

Z. SURAJ[1], J.F. PETERS[2], W. RZĄSA[3]
[1]Department of Foundations of Computer Science, Univ. of Information Technology and Management, H. Sucharskiego 2, 35-225 Rzeszów, Poland
zsuraj@wenus.wsiz.rzeszow.pl
[2] Computer Engineering, Univ. of Manitoba, Winnipeg MB R3T 5V6, Canada
jfpeters@ee.umanitoba.ca
[3]Institute of Mathematics, Pedagogical University, Rejtana 16A, 35-310 Rzeszów, Poland
wrzasa@univ.rzeszow.pl

Abstract. Decision algorithms useful in classifying meteorological volumetric radar data are the subject of described in the paper experiments. Such data come from the Radar Decision Support System database of Environment Canada [18] and concern summer storms created in this country. Some research groups used the data completed by RDSS for verifying the utility of chosen methods in volumetric storm cells classification. The paper consists of a review of experiments that were made on the data from RDSS database of Environment Canada and presents the quality of particular classifiers. The classification accuracy coefficient is used to express the quality. For three research groups that led their experiments in a similar way it was possible to compare received outputs. Experiments showed that the SVM method [14] and rough set algorithms [11],[17] which use object oriented reducts for rules generation classify volumetric storm data better then others classifiers [1],[2],[8],[9].

Keywords: rough sets, cross-validation, data mining, knowledge discovery in databases, pattern recognition.

1 Introduction

Decision algorithms useful in classifying meteorological volumetric radar data are the subject of experiments described in the paper. The volumetric radar data come from the Radar Decision Support System database of Environment Canada [18]. They concern summer storm creating in this country. The RDSS collects meteorological data by conducting a volume scan [6]. The system detects storm onset regions called storm cells. Besides, a fixed set of parameters is determined and used to characterize detected phenomenon. In case a storm is detected, values of these parameters are measured. It seems natural to try to classify storms with respect to their kind on the ground of measured parameters. Because of data incompletness, complex evolution of storm cells and high dimensionality of the data, choice of patterns which will recognize type of storm with high accuracy is a real challenge. Some research groups used the data completed by RDSS for verifying the utility of chosen methods in volumetric storm cells classification.

The paper consists of synthesis of outputs that have been received by noticed research groups, and it compares the outputs.

The structure of the paper is as follows. The next section consists of meteorological data description. Section 3 presents methods that have been used by researchers during experiments. Methodology of described experiments is a topic of Section 4 and in Section 5 outputs are presented. The last section contains conclusions.

2 Data

For 577 storm cells localized by the Vivian radar in Manitoba for summer of 1997 and by the Broadview radar for July of 1998 and collected in RDSS database of Environment Canada values of 22 parameters have been computed. Table 1 presents types of the parameters and their values. There are 4 kinds of storms distinguished: hail, rain, tornado, wind. In many cases**,** the following situation is noticed: all 22 parameters have the same values but kinds of storm are different. Table 2 presents structure of the phenomenon. The situation suggests low precision of some measurements or lack of

some important attributes. Reading the data as a decision table (with 577 objects, 22 conditions and 4 decision classes)**,** one may say it is an inconsistent table. For changing the data into a consistent table, 6 new decisions have been generated and assigned to objects that caused the inconsistent character of the table. Table 3 shows the structure of the data with 10 decision classes.

Table 1. List of derived attributes and one decision

Feature	**Data type**	**Description**
Z value	R+	Height offset [km]
Extent	(N+, N+)	Extent of a cell [km]
Core volume	N+	Volume of a cell core [km]
Core height	R+	Height of a cell core
Supercell severity	{0, 1, 2, 3}	Heuristic
Wind gust severity	{0, 1, 2, 3}	Heuristic
Hail occurence	{0, 1, 2}	Heuristic
Core tilt angle	R+	Core fitted angle
Supercell flag	{0, 1}	Heuristic
Joint count	{0, 1}	Has a cell joined?
Split count	{0, 1}	Has a cell split?
Core tilt vector	{R+, R+, R+}	Centroid parameters
Velocity set flag	{0, 1}	Next field available?
Velocity	{R+, R+, R+}	[km/h, km/h, km/h]
Core size	{N+, N+}	Size of a cell core
Orientation	R+	Orientation of a vector
Cell type	{1, 2, 3, 4}	Decision

Table 2. Original distribution of the data

Class number	**Class name**	**Number of patterns**	**Number of patterns with 2 decision classes**
1	Hail	166	16
2	Rain	54	32
3	Tornado	265	58
4	Wind	92	40

Table 3. Distribution of the data in 10 decision class table

Class number	**Class name**	**Number of patterns**
1	Hail	150
2	Rain	22
3	Tornado	207
4	Wind	52
5	Hail or rain	20
6	Hail or tornado	10
7	Hail or wind	0
8	Rain or tornado	33
9	Rain or wind	33
10	Tornado or wind	50

3. Review of methods used in experiments

The paper consists of outputs of experiments made by 4 research teams:

- M. Alexiuk, N. Pizzi W. Pedrycz [1],[2],[8],[9];
- L. Ramirez, W. Pedrycz, N. Pizzi [14];
- J. F. Peters, Z. Suraj, N. Pizzi, W. Pedrycz, S. Shan [11];
- Z. Suraj, W. Rząsa [17].

The teams used the data to verify quality of different kinds of classifiers.

M. Alexiuk, N. Pizzi W. Pedrycz [1] used data gathered by Vivian radar in Manitoba during the summer 1997. The data contained 165 storm cells and all 22 derived attributes (see Table 1). Decision has 4 classes.

Next classifying methods have been checked:

- linear discriminator using pseudo inverse of training set matrix;
- C 4.5 decision tree has been used twice – with and without gain ratio;
- neural network
- k-nearest neighbour (k-NN)
- fuzzy c-mean (FCM)

The team used also several preprocessing techniques in order to reduce the amount of input data. Principal components analysis, genetic algorithm and fuzzy interquartile encoding were used.
L. Ramirez, W. Pedrycz, N. Pizzi [14] verified classifiers by whole of the data, it is 577 objects and 22 derived attributes. They used both 4 and 10 decision class tables for their experiment.

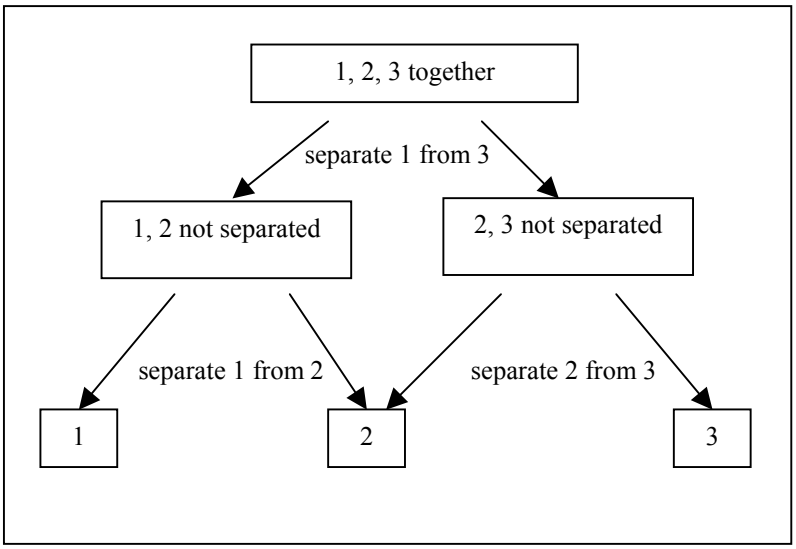
List of classifiers that have been checked is as follows:

- support vector machine (SVM)
- radial basis function (RBF)
- neural network
- k-nearest neighbour

SVM is a classifier based on optimal hyperplanes or optimal surfaces generation method. After interpretation of objects from a decision table as points of space R^n (n is a number of numeric conditions), the hyperplanes (surfaces) separating decision classes in the best way are searched. Best here means as little as possible misclassified objects and as maximal as possible distance between the separating hyperplane (surface) and the decision classes. For data with more than two decision classes couple of approaches are used:

- find a set of hyperplanes or surfaces such that each of them separates one decision class from remaining ones
- find a set of hyperplanes or surfaces that separates decision classes according to a specific graph, so called DDAG. The DDAG for 3 decision class table shows Figure 1.

Figure 1. Scheme of DDAG for 3 classes table



The RBF classifiers were implemented using the MATLAB neural networks toolbox.

Teams of J. F. Peters, Z. Suraj, N. Pizzi, W. Pedrycz, S. Shan [11] and Z. Suraj, W. Rząsa [17] used the Rosetta and the RSES systems [15], [16] to verify classifiers implemented in the systems. They are based on the rough set theory. During experiments of these teams**,** accuracy of rules has been tested. The rules have been generated with utilizing some kinds of reduct sets and decomposition tree. Moreover, Z. Suraj and W. Rząsa had repeated experiment after conditions values' discretizing.

The following reducts were generated:

- decision relative full reduct (FR) – is a set of conditional attributes that are necessary and sufficient to discernible objects from different decision classes in the same degree as by using all conditions;
- decision relative *k*-object oriented reduct (*k*-OOR) - is a set of those conditional attributes that are necessary and sufficient to discernible object *k* from all objects with another decision values. Set of such reducts for all objects is marked as OOR;
- dynamic reduct (DR) – means attributes that are a reduct of some kind (FR or *k*-OOR) on most of subtables received after resampling input table;
- genetic reduct (GR) – is a set of attributes that were selected by genetic algorithm using FR or *k*-OOR methods.

Symbol X-Y means combination of X and Y reducts.

The last team also used the data to test usefulness of the LERS system in the weather data analysis.

4 Methodology

For most of the tested classifiers it is possible to fix values of some parameters. To choose the best option researches used the data as a testing input. The process of parameters values' settling is called a preliminary testing as opposed to the part of proper experiment when accuracy of classification is tested.

4.1 Preliminary testing

For the C 4.5 method**,** in Alexiuk et al. experiments, values of condition attributes have been discretized into equal partitions over the range of each variable. On the ground of preliminary tests 6 partitions have been accepted.

In the FCM algorithm, c (maximum number of clusters) is a parameter. Experiment was led for $c = 15$ and condition attributes with real values (see Table 1).

Table 4 contains full number of parameters for the neural network method.

Table 4. Parameters for neural network method

Parameter	Value
Number of hidden layers	1
Number of hidden neurons	10 (maximum)
Number of output neurons	4
Transfer function	Hyperbolic tangent
Distance function	Manhattan distance
Training epochs	~1000
Training function	Gradient descent
Performance function	Regularized MSE
Learning rate	0.7

Ramirez et al. fixed the following parameters for their neural network algorithm after preliminary testing: 3 hidden layers with 30, 20, 15 neurons, respectively; 12500 epochs, hyperbolic tangent function is the transfer function, the initial learning rate is set to 0,01 and initial momentum term is 0,9.
In the SVM method 2 parameters are due to fix: kernel of the SVM and error penalty term *c*. The authors decided the RBF function to be the kernel. The function depends on 1 parameter γ. The data were used to fix the best values of *c* and γ by testing on them several combination of *c* and γ.
In a similar way, from classifiers based on k-NN method with the Euclidean distance and $k \in \{1, 3, 5\}$ 1-NN was chosen as the one with the best score during preliminary testing.
For all parameters associated with the generation of reducts Peters, Suraj et al. received values that the Rosetta and the RSES 1.0 systems proposed. For classifiers weight B of classification conflict resolving was received [3].
Suraj and Rząsa used the Rosetta system and the RSES 1.1 system for their experiment. After preliminary testing FR, X-FR, DR-OOR, DT [11] were aborted as a base for rules generation and in the process weights D and P were chosen.

4.2 Proper experiment

During the experiment Alexiuk et al. used 10 iterations of resampling the data over training table and test one. Presented results are average results. The size of training and test tables is not described in the paper [1].
In experiment of Ramirez et al. for both 4 decision class table and 10 decision class table the data were split into training tables and test ones in proportion 3:1. Next each training table was normalized to attain zero mean and one standard deviation. The values of the mean and the standard deviation were later used to normalize the testing tables. Columns 1 – 3 in Tables 5 and 6 show the pattern distribution in each group of training and test tables for 4 and 10 decision class tables respectively.
The same proportion of training table size and test table size was preserved by Peters et al.[11]. They randomly resampled the data 5 times. Mean distribution of patterns is presented in columns 1, 4 – 5 of Tables 5 and 6.

Table 5. Distribution of patterns for 4 decision class table in Ramirez et al. and Peters et al. experiments

Number of class	Number of training patterns [Ramirez experiment]	Number of test patterns [Ramirez experiment]	Mean number of training patterns [Peters experiment]	Mean number of test patterns [Peters experiment]
1	124	42	126	40
2	40	14	36	18
3	198	67	207	58
4	69	23	64	28

Table 6. Distribution of patterns for 10 decision class table in Ramirez et al. and Peters et al. experiments

Number of class	Number of training patterns [Ramirez experiment]	Number of test patterns [Ramirez experiment]	Mean number of training patterns [Peters experiment]	Mean number of test patterns [Peters experiment]
1	112	38	112	38
2	16	6	17	5
3	155	52	155	52
4	39	13	46	6
5	15	5	16	4
6	7	3	6	4
7	0	0	0	0
8	24	9	25	8
9	24	9	23	10
10	37	13	33	17

At the beginning of Suraj and Rząsa experiment [17] the 4 and 10 decision class tables were divided into 4 couples of training and test tables with the use of 4-fold cross-validation method. That means 3:1 proportion of training and test parts of the data in every couple. Moreover the method guaranties that mean distribution of patterns for every decision class equals theoretical distribution.
The original data showed to be too large for analysis with the use of the LERS system [5]. Therefore they have been reduced. For every couple of the training table and the test table generated at the beginning of experiment with the RSES and Rosetta systems number of attributes was decreased. From among 22 condition attributes 8-10 of them were omitted. Those were the ones with the lowest frequency in the OOR set of reducts generated for training tables. Next the data were discretized. Both processes of discretizing and OOR set determining were made with the use of the RSES system. For such prepared data the RSES, LERS and Rosetta systems were used in the analysis.

5 Results

For all experiments, the main coefficient of classifier's quality is accuracy. It is the ratio of correctly classified objects to all objects. Tables 7-11 present results of experiments for 4 teams. D (or P) letter in Tables 10 and 11 means score received with the use of D (or P) weight, respectively. The other weight gave a bit worse accuracy. Moreover, outputs of Ramirez, Pedrycz, Pizzi team and Peters, Suraj et al. and Suraj, Rząsa teams are presented on common graphs (see Figure 2 and Figure 3) because methodology of these three teams' experiments is similar and the outputs may be compared. Short view of Table 11 and last two columns of Table 10 shows that loosing 8-10 attributes didn't change the quality of classification in a considerable degree. Apart from accuracy of classification used to compare different classifiers, time of analysis made by individual classifiers was also measured by Ramirez et al., and Suraj, Rząsa teams. Because experiments were led on different computers only the rank of time may be compared. The SVM method took about 1 minute for the analysis of a couple of training and test tables, 1-NN, OOR, GR-OOR, RBF needed about 10 minutes for the same process and NNet took about 15 minutes for 10 decision class table and less than 40 minutes for 4 decision class table.

Table 7. Alexiuk, Pedrycz, Pizzi results

Classifier	Training accuracy	Test accuracy
Pseudo-inverse	0.80	0.70
k-nearest neighbour	1.00	0.41
Inductive decision tree		
Gain ratio	0.87	0.53
No gain ratio	0.83	0.50
Time averaged	0.70	0.49
File averaged	0.67	0.48

Normalized	0.81	0.70
Multi-layer perceptron		
No momentum term	0.72	0.60
With momentum	0.94	0.72
Acceleration	0.93	0.75
Adaptive learning	0.94	0.71
RPROP	0.94	0.72
Fuzzy encoding	0.92	0.76
Fuzzy c-means		
No rejection	N/a	0.75
Rejection (mean)	N/a	0.80
Rejection (minimum)	N/a	0.77

Table 8. Ramirez, Pedrycz, Pizzi results

Decision classes	**Classifier**	**Training accuracy**	**Test accuracy**
4	SVM DDAG	0.9132	0.6863
	SVM 1-v-R	0.8835	0.6726
	RBF	0.9181	0.5486
	Neural network	0.9167	0.6390
	1-NN	1.00	0.6356
10	SVM DDAG	0.9951	0.8095
	SVM 1-v-R	0.9832	0.7899
	RBF	0.9995	0.7777
	Neural network	0.9571	0.7554
	1-NN	1.00	0.7858

Table 9. Peters, Suraj, Pizzi, Pedrycz, Shan results

Decision classes	**Computer system**	**Classifier**	**Training accuracy**	**Test accuracy**
4	Rosetta	FR	0.912	0.563
		OOR	0.912	0.729
		DR-FR	0.895	0.611
		DR-OOR	0.893	0.639
		GR-OOR	0.912	0.750
	RSES	OOR	0.904	0.594
		GR-OOR	0.893	0.603
		DT	1.00	0.659
10	Rosetta	FR	1.00	0.618
		OOR	1.00	0.743
		DR-FR	0.952	0.708
		DR-OOR	0.952	0.611
		GR-OOR	0.963	0.708

10	RSES	OOR	1.00	0.751
		GR-OOR	1.00	0.751
		DT	1.00	0.820

Table 10. Suraj and Rząsa results for original data

Decision classes	Computer system	Classifier	Not discretized		Discretized	
			Training accuracy	Test accuracy	Training accuracy	Test accuracy
4	Rosetta	OOR	0.9162	0.6030	0.9162	0.6586
		GR-OOR	0.9162	0.5944	0.9162	0.6620
	RSES	OOR	0.912	0.601 D	0.894	0.631 P
		GR-OOR	0.912	0.596 D	0.894	0.631 P
10	Rosetta	OOR	1.00	0.7435	1.00	0.7834
		GR-OOR	1.00	0.7418	1.00	0.7903
	RSES	OOR	1.00	0.728 P	1.00	0.773 P
		GR-OOR	1.00	0.733 P	1.00	0.773 P

Table 11. Suraj and Rząsa results for the reduced data

Decision classes	Computer system	Classifier	Discretized	
			Training accuracy	Test accuracy
4	LERS	LEM2	0.892	0.650
	RSES	OOR	0.899 P	0.637 P
	Rosetta	OOR	0.916	0.650
10	LERS	LEM2	1.00	0.761
	RSES	OOR	1.00	0.784 P
	Rosetta	OOR	1.00	0.788

Figure 2. Comparison of accuracy for different classifiers and 4 decision class table

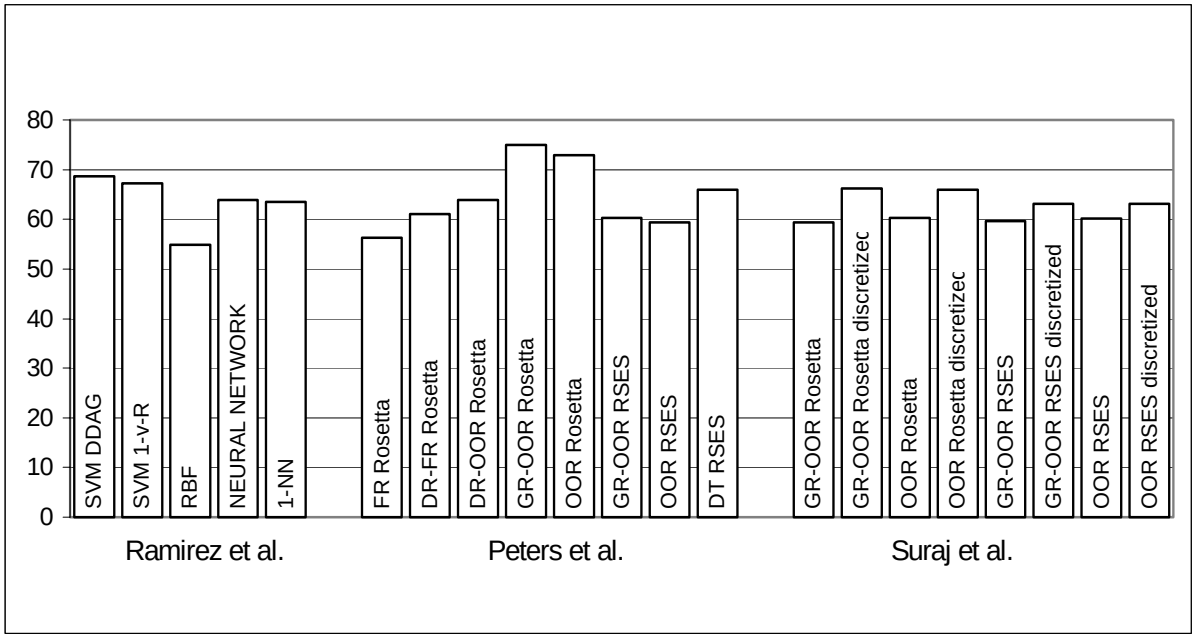
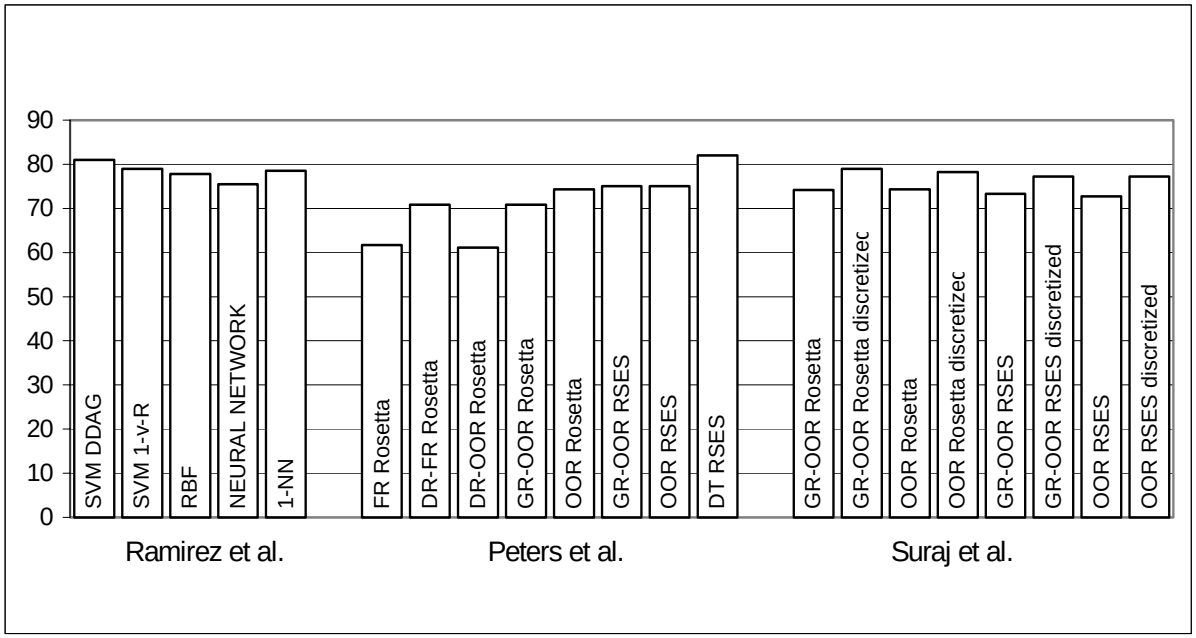


Figure 3. Comparison of accuracy for different classifiers and 10 decision class table



6 Conclusions

The best outputs for both 4 and 10 decision class tables were received by team of Peters et al. General observation corroborated in Suraj et al. experiments is that OOR and X-OOR methods are better for rule generation than the FR and X-FR methods. DT method, which gave the best result for 10 decision class table was aborted by Suraj and Rząsa in preliminary testing, because of its low quality. The 75% and 72,9% outputs for 4 decision class table are rather the effect of optimistic partition of objects in test tables. Such conclusion is inclined by results of Peters et al. for deterministic 10 decision class table and results of Suraj and Rząsa group. Connected according to formulae (0,75 * 5 + 0,59 * 4) / 9 = 0,68 for GR-OOR and (0,73 * 5 + 0,60 * 4) / 9 = 0,67 for OOR outputs of the two teams may be regarded as the outputs of 9 iterations of one experiment and be compared with mean (for 10 iterations) results of Ramirez et al. The SVM method gave a very good accuracy of classification and, moreover, it received the result in short time. Discretizing technique corroborated its utility for explored data.

Although some approaches analysed in the paper look quite promising, as demonstrated by results of described experiments, more experiments with the data are needed. Besides, it will be essential to develop methodology for classification unseen objects with using e.g. the rough integral [12], and simulated annealing method [4].

Acknowledgements. The authors of this article would like to thank all members of the teams at the Warsaw University and the Norwegian University of Science and Technology in Trondheim involved in the design and implementation of the RSES system and the Rosetta system. This research is partially supported by the National Committee for Scientific Research in Poland under grant #8T11C02519. Special thanks are due to profs. A. Skowron, J. Komorowski, and J. Grzymała-Busse for making accessible the computer systems used in our experiments.

REFERENCES

[1] Alexiuk, M., Pizzi, N., Pedrycz, W., *Classification of Volumetric Storm Cell Patterns*, Proc. of the 1999 IEEE Canadian Conference on Electrical and Computer Engineering, Edmonton, 1999.

[2] Alexiuk, M. D., *Pattern Recognition Techniques as Applied to the Classification of Convective Storm Cells*, M.Sc. Thesis. University of Manitoba, Fall 1999.

[3] Bazan, J., *A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision table*s, in Polkowski L., Skowron A. (eds): Rough Sets in Knowledge Discovery 1. Methodology and Applications, Physica-Verlag 1998, Heidelberg, 321-365.

[4] Borkowski, M., *Konstruowanie systemów decyzyjnych ze zmienną przestrzenią atrybutów*, M. Sc. Thesis, supervisor: A. Skowron, Institute of Mathematics, Warsaw University, 2000 (in Polish).

[5] Grzymała-Busse, J.W., *A new version of the rule induction system LERS*, Fundamenta Informaticae 31 (1997), 27-39.

[6] Dietrich, J., *Report on Project EC-NRC*, Spring 2000.

[7] Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A., *A Rough Set Perspective on Data and Knowledge*. In L. Polkowski and A. Skowron (Eds.). Rough Sets in Knowledge Discovery 1: Methodology and Applications. Physica-Verlag, Heidelberg, 1998.

[8] Li, P. C., Pizzi, N., Pedrycz, W., *Classification of Hail and Tornado Storm Cells Using Neural Networks*, Project EC-NRC Research Reports, 1999.

[9] Li, P.C., Pizzi, N., Pedrycz, W., Westmore, D., and Vivanco, R., *Severe Storm Cell Classification Using Derived Products Optimized by Genetic Algorithm*, Proc. of the 2000 IEEE Canadian Conference on Electrical and Computer Engineering, Halifax, 2000.

[10] Pawlak, Z., Rough sets – theoretical aspects of reasoning about data, Kluwer Academic Publ., Dordrecht 1991.

[11] Peters, J.F., Suraj, Z., Pizzi, N., Pedrycz, W., Shan, S.: *Classification of Volumetric Storm Cell Patterns Using Rough Set Methods*, in: Pattern Recognition‘2001 [to appear].

[12] Pawlak, Z., Peters, J.,F., Skowron, A., Suraj, Z., Ramanna, S., *Rough Measures: Theory and Applications*, in: Proceedings of Rough Set Theory and Granular Computing (RSTGR’2001), May 2001, Japan.

[13] Pawlak, Z., Peters, J.,F., Skowron, A., Suraj, Z., Ramanna, S., *Rough Measures: Theory and Applications* (in preparation).

[14] Ramirez, L., Pedrycz, W., Pizzi, N., Storm Cell Classification With the Use of Support Vector Machines (draft version).

[15] The ROSETTA WWW homepage, http://www.idi.ntnu.no/~aleks/rosetta/

[16] The RSES WWW homepage, http://logic.mimuw.edu.pl/~rses/

[17] Suraj, Z. Rząsa, W., *Volumetric Storm Cell Classification with the Use of Rough Set Methods*, Zeszyty Naukowe WSIiZ, Nr 1/2001 (in print).

[18] Westmore, D., *Radar Decision Support System: User Manual*, InfoMagnetics Technologies Corporation Technical Document, 1999.