Volumetric Storm Cell Classification with the Use of Rough Set Methods

Zbigniew Suraj[1], Wojciech Rząsa[2]

1. Introduction

A radar data processing system gathers meteorological volumetric radar data by conducting a volume scan. Meteorologists use these radar data to detect thunderstorms. Radar subsystem exists to allow operational meteorologists to focus their attention on the regions of interest within the volumetric radar scan known as storm cells. When a storm is found, a number of parameters are computed. There are 22 derived features and 1 decision used in the analyses such as: height offset, extent, core volume, core height, supercell severity, wind gust severity, hail occurrence, core tilt angle, supercell flag, joint count, split count, core tilt vector, velocity set flag, velocity, core size, orientation, cell type – as a decision. But it is difficult to classify detected storm cells into a specific type of storm event due to a number of confounding factors such as incomplete data, complex evolution of storm cells and high dimensionality of the data. Our objective is to identify patterns in the data that indicate, with a high degree of accuracy, the onset of a severe weather event using either the derived features of matched–cell files from the Radar Decision Support System (RDSS) database of Environment Canada [9], or the raw data of the volume scans. The classification of storm cells is a difficult problem. In this paper, the cross-validation method based on the rough set approach is used to classify storm events. We assume that the reader is acquainted with the basic notions of rough set theory [3] so all fundamental definitions are passed over. The analysis results obtained for volumetric storm cell data by using the cross-validation method are promising. In the next paper, this method will be compared with other methods, with respect to the accuracy coefficient in the classification over testing data.

2. Rough Set Classification Strategies

2.1. Introduction to classifier evaluation

In practical applications, one of the main purposes of rough set data analysis is to induce rules from data represented as information or decision systems. Then the resulting decision rules can be used to classify new and unseen objects, i.e., they can be employed to realize *classifiers* (*decision algorithms*, i.e. sets of decision rules together with methods for conflict resolving when they classify new objects). Classifiers induced from empirical data can be evaluated with respect to e.g. performance. By term performance we mean assessment of how well the classifier does in classifying new cases, according to some specified performance criterion.

A *confusion matrix C* is a $\mathrm{card}(V_d) \times \mathrm{card}(V_d)$ matrix with integer entries that summarizes the performance of an employed classifier, applied to the objects in a decision system *S*.

The entry $C(i,j)$ counts the number of objects that really belong to the decision class i, but were classified as belonging to the decision class. Of course, it is desirable for the diagonal entries to be as large as possible. Formula

[1] Chair of Computer Science Foundations, University of Information Technology and Management, H.Sucharskiego 2, 35-225 Rzeszów, Poland, zsuraj@wenus.wsiz.rzeszow.pl

[2] Institute of Mathematics, University of Rzeszów, Rejtana 16A, 35-310 Rzeszów, Poland, wrzasa@univ.rzeszow.pl

$$\frac{\sum_i C(i,i)}{\sum_i \sum_j C(i,j)}$$

defines so called *accuracy* of the classifier. The accuracy is a proportion of correctly classified objects and all objects, and it is the most popular performance measure in the machine learning literature.

2.2. Description of investigated methods for classification of unseen objects

We used the following five approaches to reducts generating and in the consequence to rules generating, used further for classification of unseen, testing objects (data). The methods for reducts generating and then decision rules are following: full reducts (FR) [3], object oriented reducts (OOR) [3], genetic reducts (GR) [10], dynamic reducts (DR) [1], decomposition tree (DT) [4].
In the following the notation *x*-*y*, where $x, y \in$ {FR, OOR, GR, DR, DT} means that we use combination of the two methods *x* and *y*, described above.

3. Methodology and Experimental Results

3.1 Methodology

Presented in the paper results of experiments have been received by applying *k*-fold cross-validation technique. That one presents as follows: data set in the form of decision table is randomly divided into *k* subsystems (with the same number of objects if possible). *k*–1 subsystems create a training data set, and the remaining subsystem is a testing data set. Received by re-sampling subtables are combined such that *k* pairs of training sets and testing sets occur. Repeating the classification process for each of the *k* pairs of training sets and testing sets we can compute the average of the estimates from each iteration to obtain an unbiased performance estimate. Each object from a data table occurs *k*-1 times into *k* training sets and exactly once is in a testing set.
Specification of experiment:
1. Data set was disjoint once, at the beginning of the experiment by cross-validation method, with $k = 5$.
2. The Rosetta system and the RSES system have been used to lead the experiment.
3. 4 classes decision table and 10 classes decision table have been independently classified. Each one has been classified twice – after discretization of data and without that.
4. Reduct generating in the Rosetta system has been made by OOR, DR-OOR, GR-OOR, FR, DR-FR methods, and in the RSES system OOR, GR-OOR, DT methods have been used.
5. The experiment has been led onto PC computer with 32 MB RAM and Pentium II processor.

3.2. Experimental results of Rosetta and RSES analysis

Results presented below have been received as the effect of Rosetta and RSES using. The Rosetta system presents results of classification as a confusion matrix. The labels of the confusion matrices are: **Actual:** the actual value of decision for a tested object, **Predicted:** the predicted value of decision by classifier for a tested object, **1:** Hail **2:** Rain **3:** Tornado **4:** Wind **5:** Hail or Rain **6:** Hail or Tornado **7** Hail or Wind **8:** Rain or Tornado **9:** Rain or Wind **10:** Tornado or Wind
Number presented at right-bottom corner of the confusion matrix means the accuracy coefficient.

		PREDICTED				
ACTUAL		**1**	**2**	**3**	**4**	
	1	19	1	6	2	67.857 %
	2	0	7	3	1	63.636 %
	3	5	1	48	3	84.210 %
	4	0	2	3	14	73.684 %
	average	79.167 %	63.636 %	80.0 %	70.0 %	76.521 %

The accuracy coefficient of the RSES system is not the same as it is for Rosetta system. Now, accuracy is a proportion of correctly classified objects and all classified objects. Because coverage is a proportion of all classified objects and all objects that's why it is necessary to multiply accuracy and coverage coefficients of the RSES system to compare the classification with classification in the Rosetta system.

	ACCURACY	**COVERAGE**
ALL CLASSES	0.775	0.852
D = 1	0.72	0.892
D = 2	0.636	1
D = 3	0.812	0.842
D = 4	0.857	0.736

3.2.1 Analysis of 4 classes decision table

The best result that has been received during Rosetta analysis for 4 classes decision tables is 76,5217%. It has appeared twice: after data discretization and OOR and GR-OOR methods of reduct generation. Those effects have been received in time of 5,83 min and of 1,93 min, respectively. The best outcome for undiscretized data is 72,1739%. It has been received by applying OOR and DR-OOR method in 0,9 min and 27 min time period, respectively. The best mean result 67,7552% has been received by applying GR-OOR method for discretized data.
During experiment precision of classifying of the objects from training tables has been tested. For all methods of reduct generation, exclude DR-OOR one, the same results have been noticed. It has not depended on data discretizing. The same outcome doesn't mean identical confusion matrices. The best result for training tables is 92,2077% and the best mean result is 91,3784%.
Classification of objects in RSES analysis that have been grouped in the same tables as during the Rosetta system analysis gave next outcomes. The best accuracy has been noted for DT method of reduct generation. For undiscretized data it received 77,6% with 0,852 coverage coefficient. It has been received in course of 16 s. Values of the same indexes (accuracy, coverage, time) for discretized data equal 75,8%, 0,848 and 14 s. DT method gave the best mean results of classification of both discretized and undiscretized data. In those two cases accuracy coefficient equals 64,1% and 70,7%, whereas mean results for OOR and GR-OOR methods of reducts generation are very similar and equal 57,2% ± 0,1% and 62,4% ± 0,1% for undiscretized and discretized data. The classification accuracy coefficient for all training tables with DT method applying equals 100% and mean cover coefficient equals 0,781 for undiscretized data and 0,78 for discretized one.

3.2.2 Analysis of 10 classes decision table

The best classification for testing tables reached 88,6956% on discretized data and 80,8695% on undiscretized ones, after both GR-OOR and OOR method of reduct generation. Those results have been received for discretized data in course of 2,5 min and 7,8 min, respectively, and for undiscretized ones 1,67 min and 0,9 min, respectively. The best mean result 80,0892% appeared on discretized data and OOR reduct generation method. The same method of reduct generation gave the best mean result on undiscretized data. The outcome is 76,4548%. Results of object classification from training tables equal 100%, exclude DR-OOR method.
The best result and the best average one for classification of objects from test tables have been reported after GR-OOR method applying for undiscretized data. The outcomes are 79,8% with 0,991 coverage coefficient and 73,7% with 0,989 cover coefficient respectively. 73,7% is also the best average result for discretized data with OOR and DT method applying. The best outcome for discretized data amounts to 77,1%. It has been received with 1,00 value of coverage coefficient and for OOR and DT method of reducts generation. Mean time of analysis with OOR and DT methods applied to undiscretized data equals 18 s and 17 s, respectively, whereas the analysis of discretized data took for OOR and DT method equals 24 s and 25 s, respectively. All objects from training tables have been correctly classified, independently from process of discretizating and from the kind of reducts generation method.

4. Conclusion

In the paper, we have presented a cross-validation method based on the rough set theory. Our main objective was to find the best techniques for classification of unseen objects connected with the weather data. General observation is that rules based on object oriented reducts and decomposition tree classify objects better than the rules based on full reducts. Discretization process improves classification by using the Rosetta system (except DR-OOR method).
Although our approach looks quite promising, as demonstrated by results of our 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 generalized rough membership function [5], and simulated annealing method [2].

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 supported in part by the National Committee for Scientific Research in Poland under grant #8T11C02519. Special thanks are due to prof. A. Skowron and prof. J. Komorowski for making accessible the computer systems mentioned above.

References

[1] Bazan, J., Skowron, A., Synak, P.: Dynamic reducts as a tool for extracting laws from decision tables. In: Proceedings of the Symposium on Methodologies for Intelligent Systems, Charlotte, NC, USA, Oct. 16-19, 1994, *Lecture Notes in Artificial Intelligence* 869, Springer, 346-355.
[2] Borkowski, M.: *Konstruowanie systemów decyzyjnych ze zmienną przestrzenią atrybutów*, Master Thesis, supervisor: A. Skowron, Institute of Mathematics, Warsaw University, 2000 (in Polish).

[3] Pawlak, Z.: *Rough sets – theoretical aspects of reasoning about data*, Kluwer Academic Publisher, Dordrecht 1991.
[4] Nguyen, S. H.: *Data regularity analysis and applications in data mining*, Ph. D. Thesis, Faculty of Mathematics, Computer Science and Mechanics, Warsaw University, Warsaw 1999.
[5] 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.
[6] Peters, J.F., Suraj, Z., Pizzi, N., Pedrycz, W., Shan, S.: Classification of Volumetric Storm Cell Patterns Using Rough Set Methods. In: *Pattern Recognition Letters 2002* [to appear].
[7] The ROSETTA WWW homepage, **http://www.idi.ntnu.no/~aleks/rosetta/**
[8] The RSES WWW homepage, **http://logic.mimuw.edu.pl/~rses/**
[9] Westmore, D.: Radar Decision Support System: User Manual, *InfoMagnetics Technologies Corporation Technical Document*, 1999.
[10] Wróblewski, J.: Finding minimal reducts using genetic algorithms. In: P.P. Wang (Ed.), *Proceedings of the International Workshop on Rough Sets Soft Computing at Second Annual Joint Conference on Information Sciences (JCIS'95)*, Wrightsville Beach, NC, Sep. 28–Oct. 10, 1995, 186-189.