http://en.wikipedia.org/wiki/Daniel_Kahneman

[Kaufmann, 1982] Kaufmann A. Introduction to the Theory of Fuzzy Sets, - Moscow.- 1982. (in Russian).

[Kuratovski, Mostowski,1967] Kuratovski K., Mostowski A. . Set theory – North Holland Publishing Company, Amsterdam.- 1967

[Newell et al, 1962] Newell A., Shaw J. C., Simon H. A. Empirical Explorations of the Logic Theory Machine: A Case Study in Heuristic. In J. Symbolic Logic.- 1962. -Volume 27, Issue 1- P. 102-103.

[Polya, 1945] Polya G. How To Solve It: A New Aspect of Mathematical Method.- Princeton, NJ: Princeton University Press. -1945.-ISBN 0-691-02356-5 ISBN 0-691-08097-6

[Петровский, 2002] Петровский А.Б. Основные понятия теории мультимножеств. – Москва: Едиториал УРСС. – 2002.- 80 с.

[Поспелов, 2001] Поспелов Д. Из истории развития нечетких множеств и мягких вычислений в России.- In Новости искусственного интеллекта. – 2001. – №2-3.

[Редько et al,2001] Редько В.Н., Брона Ю.Й., Буй Д. Б. , Поляков С.А. Реляційні бази даних: табличні алгебри та SQL-подібні мови. – Київ: Видавничий дім “Академперіодика”. – 2001- 198 с.

[Stoll, 1960] Stoll R. Sets, Logic and Axiomatic Theories. Freeman and Company, San Francisco.- 1960.

[Zadeh, 1965] Zadeh L. Fuzzy Sets. In Information and Control, 8(3). June 1965. pp. 338-53.

Authors' Information

Volodymyr Donchenko *– Professor, National Taras Shevchenko University of Kyiv. Volodymyrs'ka street, Kyiv, 03680, Ukraine; e-mail: voldon@bigmir.net.*

ROUGH SET METHODS IN ANALYSIS OF CHRONOLOGICALLY ARRANGED DATA

Piotr Romanowski

Abstract*: The paper presents results of efforts of increasing predicting events accuracy by increasing a set of attributes describing the present moment by information included in past data. There are described two experiments verifying such an approach. The experiments were carried on by the use of the RSES system, which is based on the rough sets theory. The data analyzed in the first experiment, concerning the weather, were reported at the meteorological station in Jasionka near Rzeszów from 1 April 2004 to 30 september 2005. The second experiment deals with exchange rates based on the money.pl news bulletin data (http://www.money.pl).*

Keywords*: rough sets, prediction, temporal data.*

Introduction

The intelligent analysis of data sets describing real-life problems becomes a very important issue of current research in computer science. Different kinds of data sets, as well as different types of problems that they describe, cause that there is no universal methodology nor algorithms to solve these problems. For instance, analysis of a given data set may be completely different, if we define a time order on a set of objects described by this data set, because the problem may be redefined to time dependencies. Moreover, the expectation of an analyst may be different for the same data set, up to the situation. Unknown objects classification on the basis of experience based on known objects is one of the essential issues of data exploration.

The rough set theory, presented by Z. Pawlak [Pawlak, 1981], is one of the most efficient tools in the data analysis. It is successfully used in many areas, such as expert systems, discovery and data mining or machine learning. In many cases, information is not only included in states of objects (attributes values), but in chronology of their occurrence as well. In such situations, we have to deal with temporal data. In the temporal series [Box, 1976], the time of object occurrence is treated literally, attributes' values are time functions, but it is not an essential condition for data to be treated as temporal ones. Information may be included in the sequence of object occurrence [Mannila, 1995], [Synak, 2005].

In the paper, it is assumed, that the size of the data, on the base of which the decision is predicted, does not exceed calculation capabilities of the computer system, and, there is a reserve of resources. Moreover, the analyzed data are chronological, that is why they may be treated similarly to time functions. Therefore, it seems reasonable to generate additional data, on the base of the previous objects and join them to the initial decision table, as new attributes. Such a case is analyzed in this paper, and two presented examples prove, that such an approach may increase efficiency of right decisions generating.

Next sections include a short description of basic concepts and algorithms used in the paper and the way of decision table extension. In the section Data and experiments, data, the methodology end experimental results are presented. Some and suggestions for the further research are provided in section Conclusions.

Basic Concepts and Algorithms

In the rough set theory, it is assumed, that the known world may be represented as a set of U objects [Skowron, 1993a], [Polkowski, 2002], [Bazan, 2000]. Each u_i object in U is described by a set of its attributes a_j. Object's attributes may have different meanings, and originally, they are described in different ways, but for practical purposes, they are recorded as values of certain attribute a from the finite set of values V_a, usually, the set of integer numbers (data discretization [Nguyen, 1997]). Consequently, the known world is represented by a table of numbers, rows of which are individual objects, and columns include values of corresponding attributes for individual objects. When one column is distinguished as a decision column d, the table of information is called a decision table **A**=(U,A,d), where U is a set of objects, A is a set of conditional attributes. The decision d divides a set of objects U into classes containing objects having the same value of the decision attribute. When a new object, with known values of conditional attributes and unknown value of decision appears, the decision table has to enable its classification to the certain class. Premise for such a classification is a possible similarity of the new object's conditional attribute set to conditional attributes of objects from one of the classes.

Due to big sizes of analyzed data and high complexity of deterministic algorithms there are used algorithms of lower accuracy, but of the lower complexity as well.

Decomposition trees are to divide data into fragments of the size defined earlier, which are represented as decomposition tree leaves. The global and local discretization can be used. More precise description of

classification according to decision trees and ways of discretization are presented in works [Bazan, Szczuka, 2000], [Nguyen, 1997].

Four methods of rules generating have been accepted in the paper.

In the exhaustive method the deterministic algorithm is used, that calculates all minimal rules (i.e. rules with a minimal number of descriptors on the left hand side). The discernibility matrix is generated, and on the base of it, there is build a logical formula, stating for attributes of each two objects if the objects are distinguishable. Such a formula is transformed into the simpler form by the means of absorption laws and new rules are created on the base of this formula. More precise description can be found in the work [Bazan, 2000]. The genetic method is modeled on the mechanism of genes' evolution in the nature. The initial set can be created even in random way. Then, there is a process of weaker elements elimination and better elements modification [Bazan, 2000]. The covering method is based on a set of rules generating on the basis of objects' subsets and than, creating their joint element [Bazan, 2000]. The LEM2 algorithm is based on local covering determination for individual objects from certain decision class, as presented in [Grzymała-Busse, 1997].

Decision Table Extension

Additional attributes in the 'present' object can be created on the basis of attributes' values in previous objects in many ways. The right choice should result from the knowledge of phenomena peculiarities, described by certain data. In this paper, such additional information are not used. There are analyzed the results of extensions of the decision table by sums of values of the present and previous day:

$$a_i' = a_i + a_{i-1}. \tag{1}$$

(such extensions of the decision table are special cases of the autoregression model usage for autoregression coefficients values accepted in advance [Box, 1976]),

and the results of extension data by coefficients A, B, C of parabola:

$$f(x)=Ax^2+Bx+C \tag{2}$$

which interpolates the data of three last days (see experiment 2).

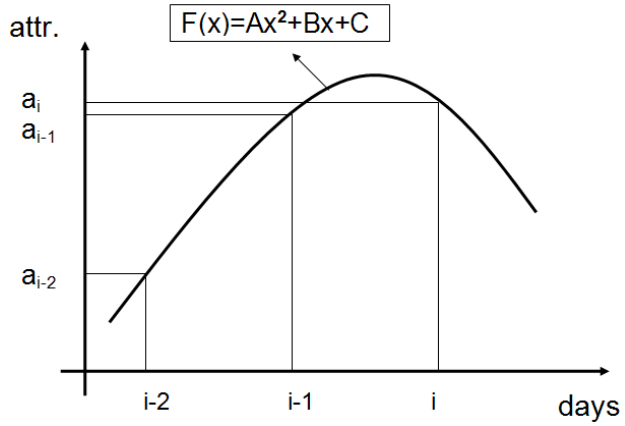

Fig. 1. Creation of new attributes A, B, C on the base of three days data

May be, it could be more profitable, to take into consideration longer preceding time, derivative equivalents or different ways of generating additional information from the past for each attribute. It could be the subject of further investigation.

Data and experiments

In all calculations, there was used cross-validation method [Michie, 1994] for data classification. Objects (rows) of the decision table are randomly divided into two sets in the ratio, that is the calculation parameter (it was accepted 4:1). The first set is the base of decision tree or rules generating, the second one is for testing [Skowron, 1993]. The accuracy of decision predicting is expressed by the means of so called "confusion matrix" [RSES, 2006], Fig. 2.

Results of experiments by cross-validation method: pogJutroT

		Predicted						
		0	2	5	1	4	No. of obj.	Accuracy
Actual	0	5.8	2.6	0.4	1.6	2.2	38.8	0.459
	2	2.4	2.8	0	0.4	1	33.6	0.438
	5	0.2	0.2	0	0.2	0	9	0
	1	1.4	0.2	0	0.8	0.2	14.6	0.217
	4	1.6	1.8	0	0	3	13	0.481
	True positive rate	0.55	0.45	0	0.21	0.45		

Total number of tested objects: 109
Total accuracy: 0.433
Total coverage: 0.264

Fig. 2. The confusion matrix

Calculations efficiency was accepted as a product of "coverage" and "accuracy' (total coverage and total accuracy). There were accepted two kinds of classifiers: decomposition trees [Bazan, 2000], [Nguyen, 1997] and decision rules [Bazan, 1998].

Six sets of results were obtained from each experiment, which illustrate the effect of primary data extension.

The profit of efficiency, resulting from the information table broadening is called a proportion:

$$\text{profit} = \frac{\text{efficiency (extended data)} - \text{efficiency (primary data)}}{\text{efficiency (primary data)}} \quad (3)$$

Experiment 1

There are analyzed the weather data, based on measurements made in the Meteorological Station in aeroplain station Jasionka near Rzeszów from1 April 2004 to 30 september 2005.

In 548 days there were measured the following data: temperature, dew, humidity, pressure, visibility, wind and cloud cover. All of them were treated as parameters of the decision table. Moreover, there were recorded such events as : fog, rain, hail, snow or storm. They were treated as decisions. Events were presented in numerical way, by accepting the following denotations: 0 – none, 1 – fog, 2 – rain, 3 – hail, 4 – snow, 5 – storm. Table 1. illustrates fragment of the weather data.

Table 1. Fragment of the weather data

temp	dew	humidity	pressure	wisibility	wind	clouds	events
[F. deg.]	[1 - 100]	[%]	[hPa]	[0 - 20]	[m/s]	[0 – 10]	[0 - 5]
43	31	58	1018.7	5	0	3	0
35	19	43	1018.6	20	10	0	0
38	15	40	1019	20	12	0	0
47	21	51	1019.2	7	10	4	2
50	38	75	1019.4	6	5	4	2

Figs 3 and 4 illustrate temperature and relative humidity (average of the day) and their changes.

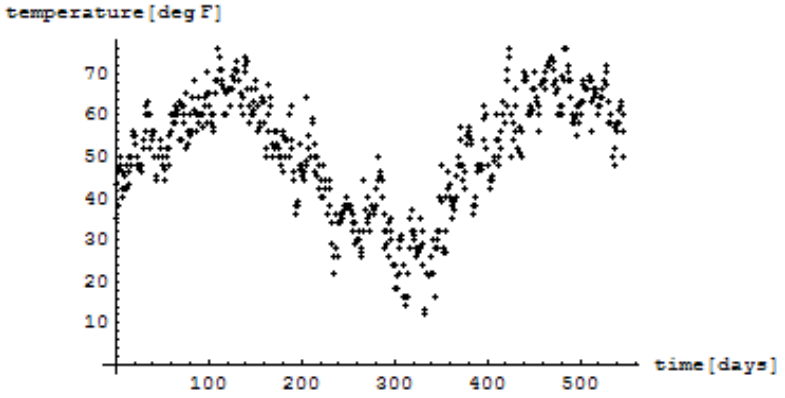

Fig. 3. The temperature (Fahrenheit's scale)

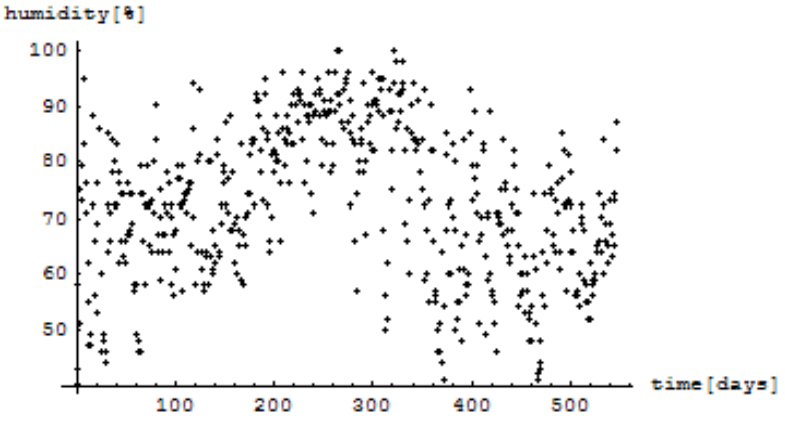

Fig. 4. Relative humidity [%]

Fig. 5 illustrates two days' sums of temperature.

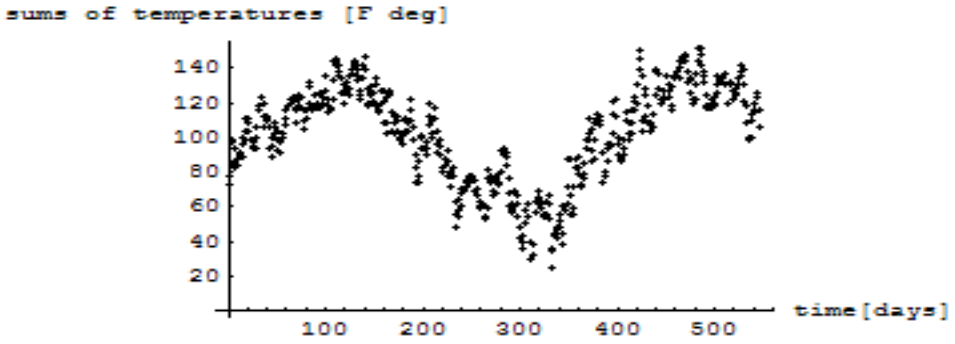

Fig. 5. Two days' sums.

The data analysis, which aim to predict phenomena – event of the preceding day, the decision 'n' is replaced with the following day event 'n+1'. Table 2 illustrates such a movement of the decision table.

Table 2. The decision table. A following day event is a decision

temp	**dew**	**humidity**	**pressure**	**wizability**	**wind**	**clouds**	**events**	**following day events**
[F. deg.]	**[1- 100]**	**[%]**	**[hPa]**	**[0 - 20]**	**[m/s]**	**[0 – 10]**	**[0 - 5]**	**[0 - 5]**
								0
43	31	58	1018.7	5	0	3	0	0
35	19	43	1018.6	20	10	0	0	0
38	15	40	1019	20	12	0	0	2
47	21	51	1019.2	7	10	4	2	2
50	38	75	1019.4	6	5	4	2	

For each of six possibilities there were performed four cycles of calculations (by the cross – validation method), the average calculation efficiency and the profit were calculated. The results are presented in Table 3.

Table 3. Experiment 1 – results

Exhaustive algorithm			**Covering algorithm**		
Data	primary	primary+ sums	Data	primary	primary+ sums
	0.347	0.404		0.130	0.249
	0.348	0.418		0.128	0.219
	0.376	0.400		0.115	0.209
	0.350	0.409		0.113	0.248
average	**0.355**	**0.408**	**average**	**0.121**	**0.231**
profit		**14.75%**	**profit**		**90.52%**

LEM 2 algorithm

Data	primary	primary+ sums
	0.113	0.125
	0.132	0.115
	0.113	0.102
	0.130	0.128
average	**0.122**	**0.117**
profit		**-3.68%**

Genetic algorithm

Data	primary	primary+ sums
	0.3520	0.4130
	0.3360	0.3433
	0.3470	0.387
	0.368	0.391
average	**0.351**	**0.384**
profit		**9.34%**

Decomposition tree (global method)

Data	primary	primary+ sums
	0.1547	0.1364
	0.1233	0.1324
	0.1547	0.114
	0.126	0.099
average	**0.140**	**0.121**
profit		**-13.69%**

Decomposition tree (local method)

Data	primary	primary+ sums
	0.1216	0.1207
	0.1483	0.1019
	0.1289	0.111
	0.149	0.118
average	**0.137**	**0.113**
profit		**-17.42%**

Although, the second experiment presents results obtained in the analysis of the data of different areas, results are quite similar.

Experiment 2

There are analyzed exchange rates of USD, Euro and CHF to PLN during 200 working days from 24 March 2007. Table 4. illustrates fragment of the Euro exchange rate data.

Table 4. Fragment of the Euro exchange rate data

USD	Euro	CHF	decision
[PLN]	[PLN]	[PLN]	(following day)
3.1553	3.9801	2.561	1
3.2012	4.0298	2.5948	-1
3.1936	4.0176	2.5888	1
3.186	4.035	2.5939	0
3.2104	4.0414	2.5945	1
3.2288	4.059	2.6027	1

Figure 6 presents Euro exchange rate.

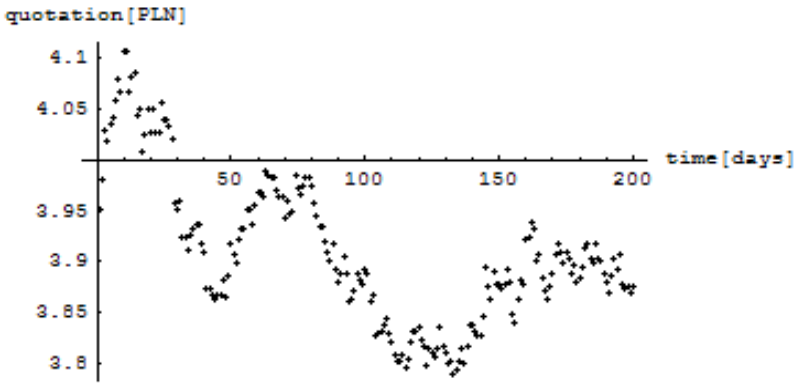

Fig. 6. Euro exchange rate

As a decision, there was accepted a EURO exchange rate in the following day in the form: decrease of more than 0.01 PLN – decision = -1, increase of more than 0.01 PLN – decision = 1 else – decision = 0. For the decision table built in such a way, the calculations were carried on in the same way as in experiment 1. The results are as follows:

Table 5. Experiment 2 – results.

Exhaustive algorithm

Data	primary	primary+ sums
	0,330	0,339
	0,333	0,313
	0,313	0,318
	0,327	0,369
average	**0,326**	**0,335**
profit		**2,84%**

Covering algorithm

Data	primary	primary+ sums
	0,114	0,275
	0,096	0,256
	0,125	0,282
	0,076	0,294
average	**0,103**	**0,277**
profit		**169,55%**

LEM 2 algorithm

Data	primary	primary+ sums
	0,126	0,144
	0,137	0,135
	0,159	0,137
	0,173	0,128
average	**0,149**	**0,136**
profit		**-8,51%**

Genetic algorithm

Data	Primary	primary+ sums
	0,301	0,349
	0,315	0,338
	0,305	0,385
	0,306	0,344
average	**0,307**	**0,354**
profit		**15,37%**

Decomposition tree

Data	primary	primary+ sums
	0,154	0,156
	0,161	0,118
	0,128	0,203
	0,142	0,133
average	**0,146**	**0,153**
profit		**4,30%**

Decomposition tree

Data	primary	primary+ sums
	0,136	0,187
	0,187	0,133
	0,118	0,171
	0,150	0,127
average	**0,148**	**0,155**
profit		**4,70%**

Exhaustive algorithm

Data	primary	primary+ ABC
	0,33	0,368
	0,333	0,368
	0,313	0,389
	0,327	0,384
average	**0,3258**	**0,3773**
profit		**15,81%**

Genetic algorithm

Data	Primary	primary+ ABC
	0,301	0,353
	0,315	0,358
	0,305	0,347
	0,306	0,358
average	**0,307**	**0,354**
profit		**15,31%**

Conclusions

Our main objective was to find the best method classifying unseen objects connected with two sets of chronologically arranged data (the weather data and the exchange rates data). The original data was extended by additional information. The experiments were executed by the use of six rough set algorithms implemented in the RSES system.

Classifications methods and theirs variants can be divided into ‘time consuming’ and ‘economical’ ones. The exhaustive and genetic methods are ‘time consuming’, whereas the others are ‘economical’.

Generally, ‘economical’ methods and theirs extensions result in the loss of decision accuracy, except for the covering method of rules generating, but even this one, in spite of significant advantages in the used extension, does not give satisfying precision, as the precision is below the random choice. Extension of data by two days sums of attributes or coefficients of parabolas, using ‘time consuming’ methods are more satisfying, as they give significant profits and according to the author of this paper, encourage to further experiments.

Bibliography

[Bazan, 1998] Bazan, J.: A comparison of dynamic and not-dynamic rough set methods for extracting laws from decision table. In L. Polkowski, A. Skowron (eds.), Rough sets in knowledge discovery, Physica-Verlag, Heidelberg, pp. 321—365 (1998)

[Bazan, 2000] Bazan, J., Nguyen, H., Nguyen, S., Synak, P. and Wróblewski, J.: Rough set algorithms In classification problem. In L. Polkowski, S. Tsumoto, and T. Lin, editors, Rough Set Method and Aplications, Physica-Verlag, Heidelberg New York , , pp. 49--88 (2000)

[Bazan, Szczuka, 2000] Bazan, J., Szczuka, M., RSES and RSESlib – A collection of tools for rough set computations. Lecture Notes in Artificial Inteligence 3066, Berlin, Heidelberg:Springer-Verlag, , pp. 592 – 601 (2000)

[Box, 1976] Box, G., Jenkins, G.: Time series analysis: Forecasting and control. Holden-Day, San Francisco, CA, 2. edition, (1976)

[Grzymała-Busse 1997] Grzymała-Busse, J.: A new version of the rule induction system LERS. Fundamenta Informaticae, vol. 31(1), pp. 27--39 (1997)

[Mannila, 1995] Mannila, H., Toivonen, H., Verkamo, A.: Discovering frequent episodes in sequences, in U. Fayyad, R. Uthurusamy. First International Conference on Knowledge Discovery and Data Mining KDD. AAAI Press, Montreal, Canada, pp. 210--215 (1995)

[Michie, 1994] Michie, D., Spiegelhalter, D, Taylor, C.: Machine learning, neural and statistical classification. Ellis Horwood, New York, (1994.)

[Nguyen, 1997] Nguyen, H.: Discretization of Real Value attributes, Boolean reasoning approach. Ph. D. thesis, supervisor B. Chlebus, Warsaw University, (1997)

[Pawlak, 1981] Pawlak, Z.: Information systems – theoretical foundations. Information systems, vol. 6, , pp. 205--218 (1981)

[Polkowski, 2002] Polkowski, L.: Rough sets: Mathematical foundations. Advances in Soft Computing. Springer-Verlag, Heidelberg, Germany, (2002)

[Skowron, 1993] Skowron, A.: Boolean reasoning for decision rules generation, in J. Komorowski, Z. Raś. Seventh International Symposium for Methodologies for Inteligent Systems ISMIS, vol. 689 Lecture Notes in Artificial Inteligence. Springer-Verlag, Trondheim, Norway, pp. 295--305. (1993)

Skowron , A.: A synthesis of decision rules: applications of discernibility matrices, Proccedings of the Conference on Intelligent Information Systems, Practical Aspects of AI, Augustów, Poland, June 7-11, pp. 30--46 (1993)

[Synak, 2005] Synak, P.: Temporal Templates and Analysis od Time Related Data, in W. Ziarko, Y. Yao (eds.), Second International Conference on Rough Sets and Current Trends in Computing, RSTC 2000, Banff, Canada, October 2000, Lecture Notes in Artificial Inteligence 2005, Springer, pp. 420--427 (2005)

[RSES, 2006] RSES Homepage http://logic.mimuw.edu.pl/~rses

Authors' Information

Piotr Romanowski *-Chair of Computer Science University of Rzeszów, Poland 35-310 ul. Dekerta 2 35 - 030 Rzeszów e-mail: proman@univ.rzeszow.pl*