AIR QUALITY MODELLING BY DECISION TREES AND HYBRID ROUGH SETS - DECISION TREES

Pavel Jirava, Miloslava Kašparová, Jiří Křupka

Univerzita Pardubice, Fakulta ekonomicko-správní, Ústav systémového inženýrství a informatiky

Abstract: *This paper deals with air quality modelling by decision trees and by hybrid rough sets-decision trees in the Czech Republic. We focused on daily observations of air polluting substances concentrations in one of the cities in the Pardubice region. After data collection, data description, and data pre-processing, we worked on the creation of classification models and the analysis of the achieved results. As modelling algorithms we selected C5.0 algorithm, boosting, and CHAID method. Finally is proposed hybrid model, where Rough Sets algorithm for the purpose of attributes reduction is used.*

Keywords: *air quality, air pollution, daily observations, model, classification, rough sets, Czech Republic*

1. Introduction

An environment is our surroundings. It includes living and non-living things around us. It is a system compounded of natural, artificial, and social components that are in interaction with one another. It is all what forms natural conditions to an existence of organisms, including human, and the preconditions of their evolution. Firstly, air, water, rocks, land, organisms, ecosystems, and energy are components of this. The weakening of components results in an imbalance and degradation of the environment.

The State environmental policy of the Czech Republic (SEP CR) [15] belongs to documents that deal with protection and quality assurance of the environment in the Czech Republic. It is a fundamental reference document for other sectors and regional policies, from the standpoint of the environment. Although SEP CR is a governmental document its implementation requires an active participation of the general public, partners in the business sector, science and research, and others. The SEP CR is a policy that should be followed by Czech Corporations, as well as other organizations, as an instrument that will assist them in their strategic and every-day operative decision-making, so as to lead not only to the creation of new economic, social, and cultural values, but also to an improvement in the quality of life and quality of the environment.

The state of the environment is regularly monitored and evaluated (annual reports of Ministry of the Environment submitted by the Government to the Chamber of Deputies of the Parliament of CR and the public) and consequently SEP CR reacts to all the important changes (negative trends) in the state of the environment. In accordance with the state of the environment transposition and implementation of European law, and the basic principles of the protection of the environment and its sustainable development, the updated SEP CR concentrates on the following four priority areas [15]:

1. Nature conservation, protection of the landscape, and biological diversity.
2. Sustainable use of natural resources, material flows, and waste management.
3. Environment and the quality of life.
4. Protection of the climate system of the Earth and prevention of long-range transport of air pollution.

This classification emphasizes not only protection of the basic components of the environment (air, water, lithosphere), but primarily integrated protection of ecosystems and the landscape (conservation of biodiversity), sustainable development, and an improvement in the quality of life. The fourth area reflects the responsibility of CR for the European and global environment (climate system, ozone layer) and the international cooperation entailed therein.

On the basis of these areas many partial goals are defined. One of the goals is to uplift air quality through defined steps and provisions. In relation to protection of human health, it is necessary to monitor the quality of drinking water and to reduce the burden on the human population resulting from the pollution of the air and foodstuffs. The Czech Hydrometeorological Institute (CHI) achieves, with the aid of various laws, the establishment and operation of a national network of monitoring stations that measure the amount of air pollution in the Czech Republic. Some of the stations in this network are designed for automated air polluting monitoring (A2PM). Measuring stations work in continuous operation and give measured values in real time to CHI centers. In the Czech Republic, 97 measuring station's A2PM work is run by CHI. Except for the results from other measuring stations outside of these 97 stations, the results are submitted in the information system. Most of the stations have analyzers to measure sulfur dioxide concentrations [SO_2], nitrogen monoxide [NO], nitrogen dioxide [NO_2], and suspended particles [PM10]. Concentrations of ozone [O_3] and carbon monoxide [CO] are only measured in few measuring stations. A selected amount of A2PM stations also measure concentrations of some volatile organic matter (benzene, toluene, xylene).

Pardubice, the seat of the Pardubice region, is situated at the confluence of the Labe and Chrudimka rivers and is one of the most beautiful towns in East Bohemia. The area of this city is practically 78 km2 and approximately 90 thousand inhabitants live there. It lies in an altitude of 215 to 237 meters above sea-level. With regard to an industrial enterprises existence, heavy traffic, and other factors, Pardubice belongs to air pollution areas.

Data used in this paper is from daily observations of air polluting substances concentrations in part of Pardubice-Dukla (Dukla) in 2007. An automated monitoring system is located in a park (in the campus of a primary school). The target of the measurement program is to evaluate the total level of concentrations and an evaluation of the effect on the population's health. Basic information about this measure is in the Table 1.

Table 1: Basic information about locality of measure

Basic Information	Value
Locality code	EPAU
Name	Pardubice - Dukla
State	Czech Republic
Owner	CHI
Basic administration unit	Pardubice
Coordinates	50° 1' 26,54 " North latitude; 15° 45' 48,78 " East longitude
Altitude	239 m
EOI - zone type	Urban
EOI - zone characteristic	Residential
Terrain	Plane, not much (sparsely) undulating terrain
Landscape	Multi-storey building (housing estates of the recent decades)
Measuring programme	Automated measuring programme

The air quality evaluation is based on the result of the weight concentrations measures of substance in the air. The evaluation of air quality by [1] is in the Table 2.

Table 2: The air quality evaluation

Air Quality	Index	SO_2	NO_2	CO	O_3	PM_{10}
		1h	1h	8h	1h	1h
		[in µg/m^3]				
Very good	1	0-25	0-25	0-1000	0-33	0-15
Good	2	25-50	25-50	1000-2000	33-65	15-30
Favorable	3	50-120	50-100	2000-4000	65-120	30-50
Satisfactory	4	120-250	100-200	4000-10000	120-180	50-70
Bad	5	250-500	200-400	10000-30000	180-240	70-150
Very bad	6	500-	400-	30000-	240-	150-

This evaluation takes the possible influence of human health into account [15]. New limits of monitoring and air quality evaluation are specified in the regulation of the Czech Republic government No: 597/2006 Coll. These limits are set separately for health protection and vegetation and ecosystems protection.

2. Problem Formulation

The goal of this paper is to create a model of air quality in a given locality through the use of selected methods. It means to design and verify a classification model through the usage of decision trees. The following are the steps of realization:

- data description and data pre-processing
- classification model creation by decision trees
- testing of classifiers and comparison of results

2.1 Data Description and Data Pre-processing

Original data was obtained from the daily observation of air polluting substances concentrations in 2007 in Dukla. In this first step we realized data cleaning, standardization, and correlation.

Data cleaning techniques [13] fill in missing values, smooth noisy data, identify outliers, and correct inconsistencies in the data. Methods used for dealing with missing values include: ignoring the objects, filling in the missing value manually, using the attribute mean to fill in the missing value, etc. [3, 5, 17]. In our case we ignored objects with missing values. The attribute means that using the most probable value or most frequent value is a convenient method in this data. Original data matrix included 365 observations. After an elimination of missing values, 330 daily observations (objects, data) described by 11 attributes (variables) were achieved. It means, we achieved data matrix O in dimension 330×11. Every observation o_i for $i = 1, 2, \ldots, 330$ can be described by the following vector $\boldsymbol{o}_i = (x_{i1}, x_{i2}, \ldots, x_{i11})$. Basic descriptive characteristics of attributes are in the Table 3. Although the air pollution rate is the result of many factors, the classification model is created on the basis of this available data.

In the determination of air quality in Dukla locality, (output attribute y_k) on the basis of the achieved data, the techniques for the air quality evaluation in Table 2 were used. It means we work with the index (class) of air quality evaluation y_k for $k = 1, 2, 3, 4, 5, 6$ and the final vector is the following: $\boldsymbol{o}_i = (x_{i1}, x_{i2}, \ldots, x_{i11}, y_k)$.

The mean monthly values (Table 3) of variables x_1, x_2, x_3, x_4 and x_5 from Table 2 measured in Dukla locality in 2007 are in Fig.1 and Fig.2. Other variables x_6, x_7 and x_8 are in Fig.3.

Table 3: Basic descriptive characteristic of input attributes

M At.	Name of Attribute	Min	Max	Mean	Std. Dev
Air polluting substances					
xx_1	Sulfur dioxide (SO_2)	1.40	113.30	7.95	7.88
xx_2	Nitrogen dioxide (NO_2)	6.00	50.20	19.90	8.03
xx_3	Carbon monoxide (CO)	128.60	1490.4	532.09	321.50
xx_4	Ozone (O_3)	9.00	105.20	51.23	22.71
xx_5	Suspended particles (PM_{10})	6.00	91.40	26.37	15.18
xx_6	Nitrogen monoxine (NO)	0.70	66.40	7.53	9.68
xx_7	Suspended particles ($PM_{2,5}$)	3.50	61.40	18.16	10.88
xx_8	Nitrogen oxides (NO_x)	7.70	168.80	32.36	24.27
Meteorological attributes					
xx_9	Solar radiation	7.10	423.80	156.42	113.17
xx_{10}	Temperature two meters above the surface of the Earth	266.10	297.70	283.16	7.48
xx_{11}	Relative air humidity	63.40	82.40	76.80	3.49

Representation of data by the index (class) of air quality evaluation y_k is in Fig.4. We can see that this locality belongs to areas with good (48.79 %) and favorable (44.55 %) air quality.

Frequently, original data should be transformed into new forms in order to perform the mining task. In our case we realized data standardization by standard deviation [3, 17].

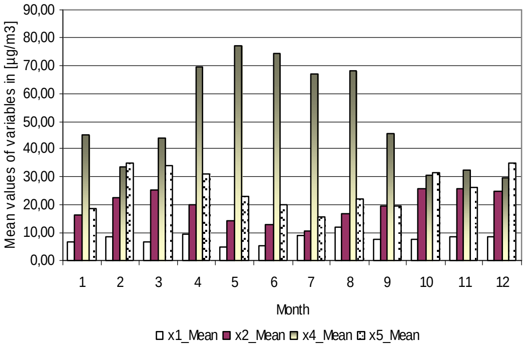

Fig.1: Mean monthly values of x_1, x_2, x_4 and x_5

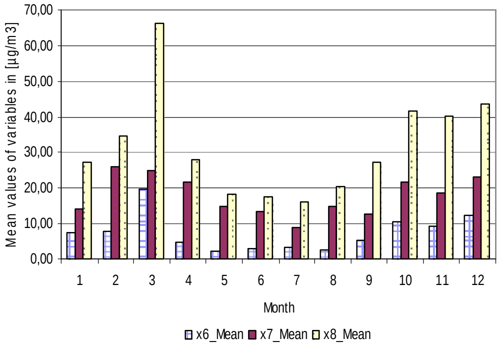

Fig.3: Mean monthly values of x_6, x_7 and x_8

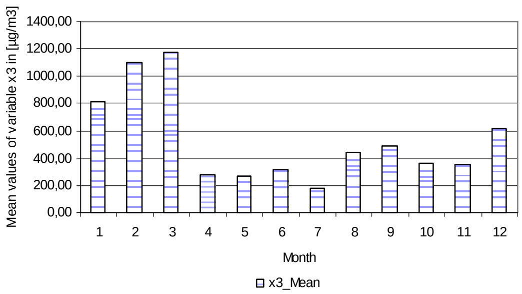

Fig.2: Mean monthly values of x_3 variable

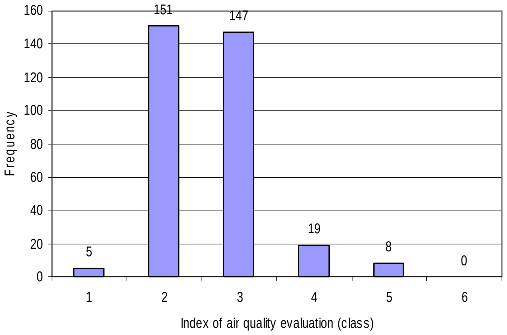

Fig.4: Representation of data by the index (class) of air quality evaluation

In the measure of the relationship between variables we used the correlation [3, 17]. The most widely-used type of a correlation coefficient is Pearson correlation coefficient ρ_{ij}. In the data matrix for classification model with 330 observations described by 11 inputs variable and 1 output variable, the top correlation was found between variables x_5 and x_7 (ρ_{ij} = 0.973) and x_6 and x_8 (ρ_{ij} = 0.975). On the basis of variables in Table 2 that are used for air quality evaluation, we eliminated attributes x_7 and x_8.

2.2 Rough Sets and Decision Trees

We propose to use in this paper Rough sets theory for prior selection of attributes from data collected in information tables. The reason is, that the classification model based only on decision trees may produce an attribute that is individually quality, but when it is selected and used to construct a complete tree, the input data with those attributes may result in nonsensical outputs and produce an inferior decision tree with poor classification knowledge.

The main goal of the rough sets (RSs) analysis is to synthesize approximation of concepts from the acquired data [7,10]. Every object we explore we associate with some information (data). Objects characterized by the same data are indiscernible in view of the available information about them. The indiscernibility relation generated in this way is the mathematical basis of RSs theory [11]. More about RSs we can found in our previous papers [4,5,6] We want to combine the utility of both Rough Sets and Decision Tree induction algorithms. This idea is simply described (more precisely is elaborated in section 3.1, and 4) in Fig.5.

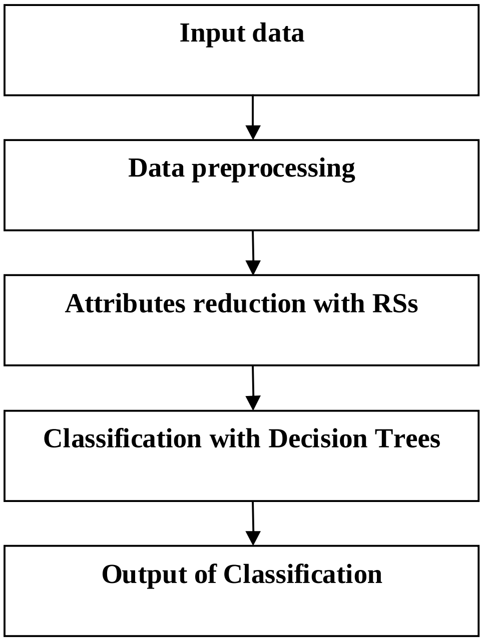
Input data
Data preprocessing
Attributes reduction with RSs
Classification with Decision Trees
Output of Classification

Fig.5: Rough - Decision Trees model

3. Classification Model Creation

The content of this paper is to describe the designed classification model (classifier) and the achieved results of classification.

For the modelling of air quality we used a data set that contains 1 dependent variable y_k and 9 independent variables x_1, x_2, x_3, x_4, x_5,x_6, x_9, x_{10} and x_{11}.

We randomly partitioned the dataset into two parts. In regards to the classification model creation, two thirds of the original dataset was allocated to the training set and the remaining objects were allocated to the testing set. Using the same objects to train and estimate their accuracy may result in misleading estimates due to overfitting. In this case if we used training set for testing we can only determine the resubstitution error R_c [3, 17]. It is the error rate in the training data set. It is calculated by resubstituting the training instances into a classifier that was constructed from them. Although it is not a reliable predictor of the true error rate on new data, it is nevertheless often useful to know.

We dealt with decision trees. The C5.0 algorithm and boosting were used in this example and focused to CHAID algorithm as well. For modelling purposes we used software Clementine Desktop 10.01. The classification model design is in the Fig.6.

3.1 The C5.0 Method and Boosting

A decision tree [13, 14] is an analytical tool. It allows developing classification systems that predict or classify future observations based on a set of decision rules. The decision tree is a straightforward description of the splits found by the algorithm. It consists of a root, nodes, and branches and leafs (terminals). Algorithms as ID3, C4.5, QUEST, etc. are used for the building of decision trees (more examples in [2, 3, 12, 13, 14, 16]).

A C5.0 method [14] works by splitting the sample based on the attribute that provides the maximum information gain [3, 5, 17]. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned.

Boosting (also Adaptive Resampling and Combinating) is a general method for improving the performance of any learning algorithm [8]. It works by building multiple models in a sequence. The first model is built in the usual way. Then, a second model is built in such a way that it focuses especially on the records that were misclassified by the first model. Then a third model is built to focus on the second model's errors, and so on. Finally, cases are classified by applying the whole set of models to them, using a weighted voting procedure to combine the separate predictions into one overall prediction. Boosting can significantly improve the accuracy of a C5.0 model, but it also requires longer training (more examples in [3, 5, 17]).

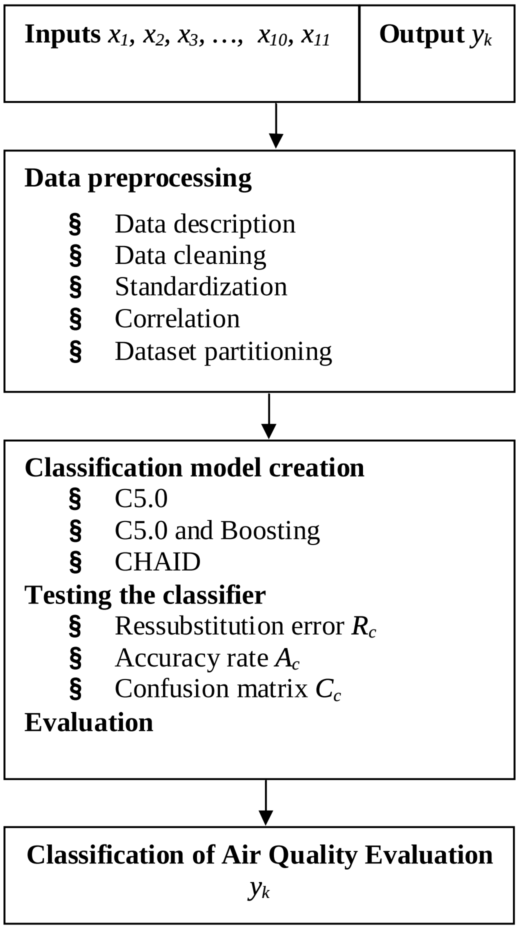

Fig.6: The classification model design

3.2 The CHAID Method

CHAID (Chi-squared Automatic Interaction Detection) is a classification method for building decision trees by using chi-square statistics to identify optimal splits. It was originally designed to handle nominal attributes only [8].

CHAID first examines the cross tabulations between each of the predictor variables and the outcome, and tests for significance using a chi-square independence test. If more than one of these relations is statistically significant, CHAID will select the predictor that is the most significant (smallest p value). If a predictor has more than two categories, these are compared, and categories that show no differences in the outcome are collapsed together. This is done by successively joining the pair of categories showing the least significant difference. This category-merging process stops when all remaining categories differ at the specified testing level. For set predictors any categories can be merged, for an ordinal set only contiguous categories can be merged [14] (more examples in [5, 12]).

3.3 Experimental Results of Classification

The resulting classifiers were tested on the train and test sets, and many tests were realized. We used the resubstitution error R_c, the accuracy rate A_c, and confusion matrix C_c [3, 5, 17], a convenient tool for analyzing the performance of a classifier. It is a square matrix that specifies the accuracy of the classifier to the classification problem. A good classifier should have a diagonal confusion matrix (all off-diagonal values are zero) [8].

The accuracy of a classifier A_c on a given test set is the percentage of test set objects that are correctly classified by the classifier. It refers to the ability of a classifier to correctly predict the class label of new or previously unseen data. The associated class label of each test object is compared with the learned classifier's class prediction for that object [3].

Achieved results by these methods are in Table 4. The mean results of tests can be seen in the Table 4. We can say that the best method of classification under the term of this problem is C5.0 with $A_{c(c5.0)}$ = 94.2 % of correct classification, with boosting it is $A_{c(c5.0\ boost)}$ = 94,43 % ($A_{c(C5.0)} < A_{c(C5.0\ boost)}$). The best result was achieved by C5.0 with boosting (99.06 % of correct classification) and the worst was CHAID method with result 88.29 % of correct classification. These results can be seen in Table 5 and Table 6.

Generally, on the basis of realized tests it means:

$$\mathbf{A_{c(CHAID)} < A_{c(C5.0)} < A_{c(C5.0\ boost)}}\ . \qquad \textbf{(1)}$$

Table 4: Mean values of tests in test data

Method	Accuracy rate A_c [in %]	Resubstitution error R_c [in %]
C5.0	94.20	94.48
C5.0 boosting	94.43	93.36
CHAID	91.41	91.67

Table 5: The best values of tests in test data

Method	Accuracy rate A_c [in %]
C5.0	96.23
C5.0 boosting	99.06
CHAID	94.44

Table 6: The worst values of tests in test data

Method	Accuracy rate A_c [in %]
C5.0	91.75
C5.0 boosting	90.4
CHAID	88.29

An example of the confusion matrix $C_{c(C5.0)}$ for classifier based on C5.0 algorithm is in Table 7. The accuracy rate is 95.12 %. The rows represent actual observed values, and the columns represent predicted values. The cell in the table indicates the number of records for each combination of final and actual values.

Table 7: Example of confusion matrix

			Final values of classifier $C_{c(C5.0)}$				
			Class				
Actual observed values	Class		1	2	3	4	5
		1	1	1	0	0	0
		2	0	55	4	0	0
		3	0	1	51	0	0
		4	0	0	0	8	0
		5	0	0	0	0	2

The result comparison of methods is in Fig.7. We can see every method achieves approximately similar results of classification.

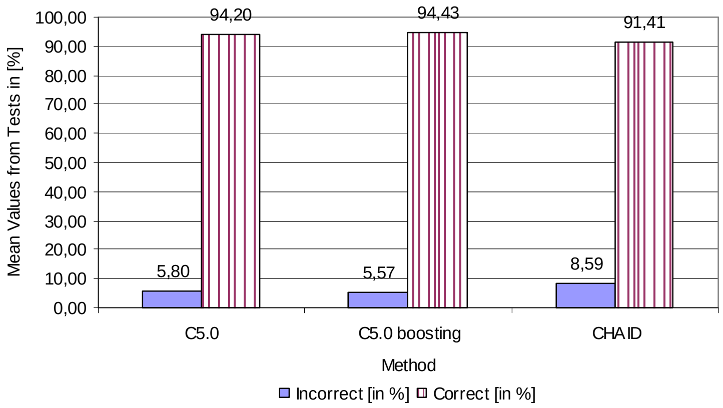

Fig.7: Results comparison of methods

4. Hybrid Rough Sets - Decision Trees Model

Creation of this model is very similar to previous classification model described in section 4. For the building of decision trees are used also algorithms C5.0, C5.0 - Boosting and CHAID. The main difference is in inclusion of Rough Sets theory algorithms in this model for the purpose of attributes reduction. Attributes reduction should improve performance of this model, decrease the size of the hypothesis space, and allow classification algorithm to operate faster and more efficiently.

Design of this classification model is in the Fig.8. On testing and verification of this model will be focused our future investigation. We want to use the same dataset like in section 3 in this paper and compare outputs from models described in Fig.6 and Fig.8. We suppose more efficient and accurate will be hybrid approach.

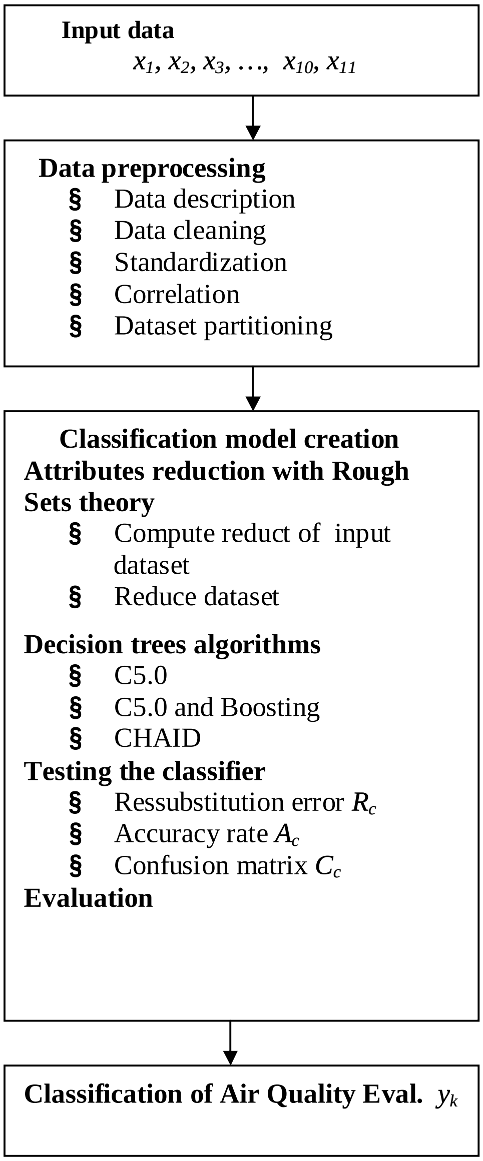
Input data
$x_1, x_2, x_3, \ldots, x_{10}, x_{11}$
Data preprocessing
§ Data description
§ Data cleaning
§ Standardization
§ Correlation
§ Dataset partitioning
Classification model creation
Attributes reduction with Rough Sets theory
§ Compute reduct of input dataset
§ Reduce dataset
Decision trees algorithms
§ C5.0
§ C5.0 and Boosting
§ CHAID
Testing the classifier
§ Ressubstitution error R_c
§ Accuracy rate A_c
§ Confusion matrix C_c
Evaluation
Classification of Air Quality Eval. y_k

Fig.8: The hybrid classification model design

5. Conclusions

Not only in the Czech Republic is air pollution and corresponding air quality belongs to very important and actual questions.

We focused on the air quality modelling in Pardubice-Dukla locality. We collected daily observations of air polluting substances concentrations described by eleven attributes and we analyzed them. In the data pre-processing step we standardized data and used correlation. On the basis of result of correlation we eliminated two attributes x_7 and x_8.

We defined output variable y_k on the basis of air quality evaluation (Table 2). In this step it is possible to use other ways of the output definition, for example to use cluster analysis, neural networks, etc. These approaches are solved for examples in [4, 9].

For the classification model creation we used decision trees. We focused on the C5.0 algorithm, boosting, and CHAID. Afterwards we analyzed and compared achieved results of classification. We state that the used methods give very similar results.

The best is the C5.0 algorithm with boosting. Achieved accuracy rate is 94.43 %. Generally class 2 (good) and class 3 (favourable) belong to the most frequent classis.

For the improvement of models it seems appropriate to use more variables that cause air pollution and work with more daily observations.

Future work: The areas where future investigations could be directed can be divided into two groups. First, it is testing and verification of hybrid Rough sets- decision trees model. Secondly it is searching of new approaches leading to model improvement and enhancement of classification accuracy.

References:

[1] Czech Hydrometeorological Institute [online], URL http://www.chmi.cz, (in Czech).

[2] GUIDICI, P. *Applied Data Mining: Statistical Methods for Business and Industry*, West Sussex: Wiley, 2003.

[3] HAN, J., KAMBER, M. *Data Mining: Concepts and Techniques*, Morgan Kaufmann Press, 2001.

[4] JIRAVA, P., KŘUPKA, J. Classification Model based on Rough and Fuzzy Sets Theory, WSEAS Computational Intelligence, *Man-Machine Systems and Cybernetic,* 2007, pp. 199-203.

[5] JIRAVA, P., KŘUPKA, J. Generation of Decision Rules from Nondeterministic Decision Table based on Rough Sets Theory. In: *Proceedings of the 4th International Conference on Information Systems and Technology Management 4°CONTECSI2007.* Sao Paolo, Brasil, s.566-573, 2007. ISBN 978-85-99693-02-5.

[6] JIRAVA, P., KŘUPKA, J. Modelling of Rough-Fuzzy Classifier. In: *WSEAS Transaction on System.* Issue 3, Volume 7, March 2008.

[7] KOMOROWSKI, J., POLKOWSKI, L., SKOWRON, A. Rough sets: a tutorial. In: *S.K. Pal and A. Skowron, editors, Rough-Fuzzy Hybridization: A New Method for Decision Making*, Springer-Verlag, Singapore, 1998.

[8] MAIMON, O., ROKACH, L. *Decomposition Metodology for Knowledge Discovery and Data Mining,* World Scientific Publishing, 2005.

[9] OLEJ, V., HÁJEK, P., KŘUPKA, J., OBRŠÁLOVÁ, I. Air Duality Modelling by Kohonen's Neural Networks, *WSEAS Environmental Science, Ecosystems & Development,* 2007, pp. 221-226.

[10] PAWLAK, Z. Rough sets. In: *Int. J. of Information and Computer Sciences*, 11, 5. 1982, pp.341-356.

[11] PAWLAK, Z. Rough set elements. In: *Rough Sets in Knowledge Discovery I – Methodology and Applications*, Physica Verlag, Heidelberg, 1998, pp.10-31.

[12] PYLE, D. *Business Modeling and Data Mining*, Morgan Kaufmann Publishers, 2003.

[13] RUSELL, S. J., NORVIG, P. *Artificial Intelligence: A Modern Approach,* Prentice Hall, 2002.

[14] SPSS Inc. *Clementine® 7.0 User's Guide*, 2002.

[15] SP: *State Policy of Environment in the Czech Republic 2004 – 2010,* Praha: Ministry of Environment, 2004, (in Czech).

[16] TURBAN, E. et al. *Decision Support Systems and Inteligent Systems,* Prentice Hall, 2005.

[17] WITTEN, I. H., FRANK, E. *Data Mining: Practical Machina Learning Tools and Techniques,* Morgan Kaufman, 2005.

Kontaktní adresy:

Ing. Pavel Jirava, Ph.D.
Ústav systémového inženýrství a informatiky
Fakulta ekonomicko-správní
Univerzita Pardubice
Studentská 95
532 10 Pardubice
E-mail: Pavel.Jirava@upce.cz
tel.: 466036001

Ing. Miloslava Kašparová, Ph.D.
Ústav systémového inženýrství a informatiky
Fakulta ekonomicko-správní
Univerzita Pardubice
Studentská 95
532 10 Pardubice
E-mail: Miloslava.Kasparova@upce.cz
tel.: 466036245

doc. Ing. Jiří Křupka, CSc.
Ústav systémového inženýrství a informatiky
Fakulta ekonomicko-správní
Univerzita Pardubice
Studentská 95
532 10 Pardubice
E-mail: Jiri.Krupka@upce.cz
tel.: 466036144

This work was supported by the scientific research project of Ministry of Environment, Czech Republic under Grant No: SP/4i2/60/07 with title Indicators for Valuation and Modeling of Interactions among Environment, Economics and Social Relations.