The data mining ensemble approach to river flow predictions

M. Cisty, J. Bezak

***Abstract*—**This paper deals with an application of data-driven ensemble methods while solving an important hydrology task - short-term flow predictions. Flood warnings several days in advance could provide civil protection authorities and the public with the necessary preparation time and could reduce the socio-economic impacts of flooding. The authors have focused on the application of an ensemble learning methodology for flow predictions, with the aim of refining the precision of the results of such modelling. Moreover, the authors demonstrate the usefulness of various steps in the data mining process, which are formalised in the so-called CRISP-DM process. They emphasize that all these steps are equally important in the modelling process and not merely in the final data-driven computations. The authors demonstrate selected methods for data pre-processing in the field of hydrology. A comparison of the ensemble modelling approach with a single model application reveals the advantage of the proposed ensemble approach. The paper describes river flow predictions in the Kysuca River watershed in Slovakia. The results from both approaches are evaluated with the help of the hydrological data which were observed in this region. An evaluation of the proposed methods shows their usefulness in river flow predictions.

***Keywords*—**data mining, ensemble models, river flow prediction

I. Introduction

THIS study deals with the application of data-driven modelling and data mining in hydrology. Data mining is an information extraction activity, the goal of which is to discover the hidden knowledge contained in (usually large) databases. Because hydrology is a very data-intensive activity, data mining has attracted increased interest in recent years [1], [2]. Various data mining tasks or tools can be used in the domain of hydrology, e.g., classification, regression, clustering, etc. When a data scientist is dealing with one of these tasks, he should go through some common process for the application of these tasks, which consists of various steps; it is usually possible to speak about six phases, which were established as the so-called CRISP-DM process (Cross Industry Standard Process for Data Mining) [3]. CRISP-DM consists of the following six phases, which are intended to be as a cyclical process (see Fig. 1).

"*The problem understanding"* is the first and most important phase. "The problem understanding" includes determining the objectives, assessing the situation, determining the data mining goals, and producing the project plan.

"*The data understanding*" phase contains a collection of the initial data into various formats, e.g., tables, the description of the data, e.g., by various statistics or graphs, the exploration of the data, the verification of the data's quality, etc.

"*The data preparation"* is a phase where construction of the datasets to be used in the modelling is accomplished. This phase includes selection of the data; in cases of using numerical data, this phase could include the selection of useful rows and columns (variables) from any available databases. The first process is the so-called sampling, while the second one is the feature or variable selection. These are followed by constructing new variables on the basis of the known ones, cleaning the data, making decisions about dealing with any missing data and, finally, constructing various data sets – e.g., for training, testing or validation purposes, and formatting these data into the form required by the software to be used.

"*The modelling phase"* includes the selection of a modelling technique, the selection of the methods for the proper tuning of these methods, and building the models which are able to address the modelling aims. In the past it was usual to search for a model optimized in some way, e.g., to find the "best" model. Nowadays, it is accepted in the hydrology modelling community that there is no one best model which is superior under all circumstances [4]. In the following study an ensemble methodology is applied to streamflow predictions. An efficient ensemble should be composed of predictors that are not only sufficiently accurate, but are also dissimilar, in the sense that the predictor errors occur in different regions of the input space [5]-[7]. Obviously, combining several identical models results in no gain. The diversity of the individual basic learners which form the ensemble in this study is achieved through the application of different learning algorithms. Each of these algorithms provides a different means for traversing the error surface; therefore, using the different training algorithms yields models that generalize differently, since they achieve different global and local minima and therefore produce different predictions [8], [9].

"*The evaluation phase"* thoroughly evaluates the model's outputs and ascertains if it properly achieves the objectives. The proper measures of the modelling quality should be selected. These includes an evaluation of the results, the

M. Cisty is with the Department of Land and Water Resources Management, Faculty of Civil Engineering, Slovak University of Technology, Bratislava, Slovakia (phone: +421259274628, e-mail: milan.cisty@stuba.sk)

J. Bezak is with the Department of Land and Water Resources Management, Faculty of Civil Engineering, Slovak University of Technology, Bratislava, Slovakia (e-mail: juraj.bezak@stuba.sk)

review process, and the determination of the next steps (back to the objectives or stepping forward to deployment as was said and as is described in Fig. 1; it is a cyclical process).

"The deployment phase" tries to use this approach for practical usage. It includes the establishment of some plan for deployment, monitoring and maintenance, and it is customary to produce a final report in which the data mining project is described and reviewed.

An application of the CRISP-DM process for a selected hydrology problem, i.e., short-term flow predictions, is presented in this paper. The structure of the paper's chapters follows the structure of the CRISP-DM process for a better description of the data mining process and the utilization of its methods in hydrology in the following parts of the paper. The methods mentioned will be applied to the Kysuca River basin in northwest Slovakia.

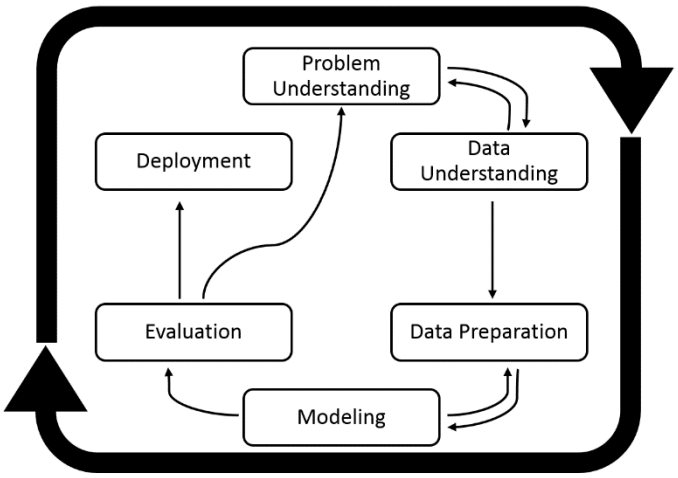


Fig. 1 CRISP-DM data mining process

II. Problem Understanding

The watershed of the Kysuca River was chosen for this study. A so-called flood warning system, in which the flow predictions are extremely important, has been established here. This watershed falls within the Vah River basin, which is a sub-basin of the Danube River. This watershed is extremely sensitive from the point of view of floods. The reason for this is that it is characterized by very quick runoff (it is located in a mountainous area). It has an irregular fan-shaped configuration, as can be seen in Fig. 2. The Kysuca watershed has an area of almost 1000 km^2.

The "time of concentration" in the Kysuca River watershed is about 2-3 hours. The time of concentration is a concept used in hydrology to measure the response of a watershed to a rain event. It is defined as the time needed for water to flow from the most remote point in a watershed to the watershed outlet [4]. It is a function of the topography, geology, and land use within a watershed. We will use this characteristic later to specify the backward time lags, from which flows will be considered as inputs for the computation of the predicted flows.

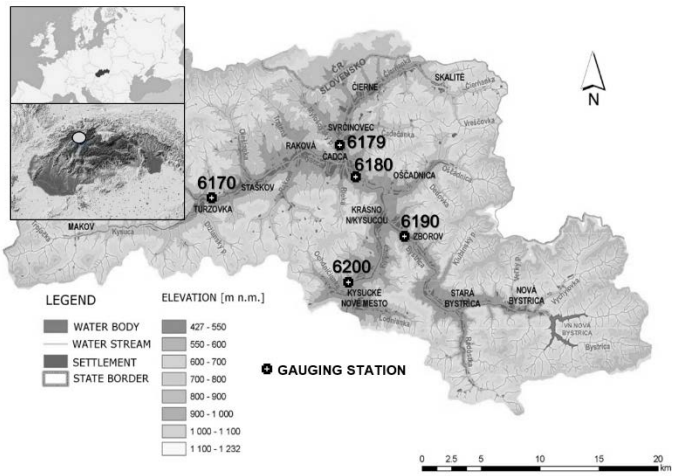


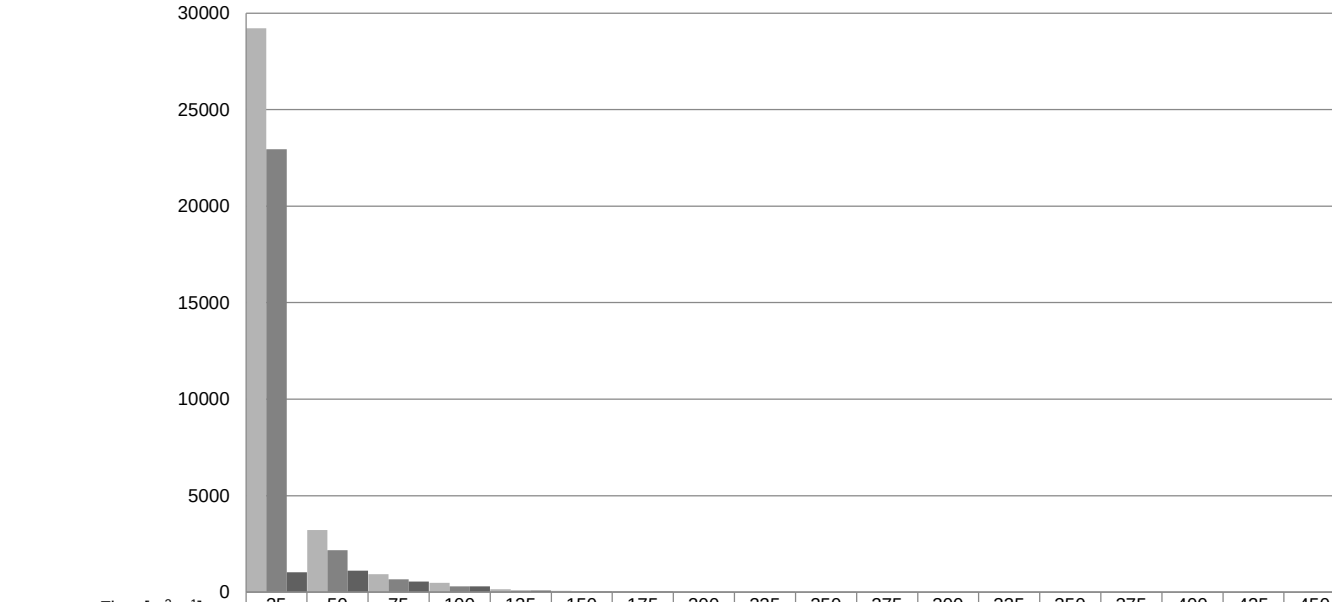
Fig. 2 Map of the area studied within the Kysuca River Basin

Various time levels of flood predictions can be considered, for example, short-term predictions (usually hours) or long-term predictions (days or weeks). In this paper we used data mining methods for short-term flood predictions. The river flow seven hours ahead of the downmost gauging station (Kysucke Nove Mesto) was chosen for the predictions. This time interval was selected on the basis of the idea that it is quite an adequate time for preparing the inhabitants of this area in the case of flood risk.

III. Data Understanding

In the study presented the above-mentioned hydrological system was represented and described by data from five measuring stations, the locations of which can be seen in Fig. 2 (dots). Three of these stations are located on the Kysuca River itself (Turzovka - 6170, Cadca–Kysuca - 6180, and Kysucke Nove Mesto - 6200), and two of them are on left-hand inflows to the Kysuca River (Zborov - 6190 and Ciernanka - 6179). The gauging stations record water levels, precipitation, and water and air temperatures. The flows are computed with the help of an appropriate rating curve for the transposition of the water level data into the flow data. The rating curve is a site-specific graphic or mathematical relationship between the water level (stage) and flow, and it is used to convert a water level record into a flow record. In this study only flows were used for the computations, because other data were erroneous or not suitable for this task. No precipitation was used, since, as can be seen in Fig. 2, the gauging stations do not reliably cover the whole watershed area; therefore, many significant storms could potentially not be registered by this gauging system. On the other hand, the flows naturally "register" every storm, because they are responses to them, so a higher degree of precision with respect to a description of the hydrological processes could be assumed in flow versus precipitation data. If a period ahead that is longer than seven hours would be chosen for the predictions, precipitation (and probably also predicted precipitation) should be used as inputs.

The observation period from which the data was collected



Flow [$m^3.s^{-1}$]	25	50	75	100	125	150	175	200	225	250	275	300	325	350	375	400	425	450
All data	29211	3228	932	485	157	74	63	32	18	22	18	16	22	14	5	6	4	1
Original Training dataset	22950	2177	671	308	86	37	34	12	4	4	2	2	3	2	2	2	2	1
Sampling Training dataset	1039	1114	556	302	83	37	34	12	4	4	2	2	3	2	2	2	2	1

Fig. 3 Histogram of the flows in the Kysuce River – Kysucke Nove Mesto gauging station

spans an interval from 1.1.2007 to 31.12.2010. In this study the authors took the flows of the Kysuca River at the Kysucke Nove Mesto gauging station as the searched output and the rest of the data were considered as the inputs. The source data set included 5 columns (5 gauging stations) and about 35,000 rows (days).

During a certain period of time in January 2010, some of the gauging stations were corrupted, so this whole month was excluded from the data set. Also, other missing values were detected in the original data set; this problem will be addressed in the next step of the data mining process, which will be described in the following chapter.

IV. Data Preparation

After the exploration of the data set, the year 2010 was chosen as the testing period. The other years were used for training the models. This decision was made, because the year 2010 was the last year in the available dataset and also because in the year 2010, large flows occurred. In flood prediction it is more important to predict high flows correctly, so 2010 was a suitable year for the testing. That is the reason why the whole data set was divided into training and testing datasets, according to Table I.

The basic input variables (the flows from all five measuring stations) are used for the purpose of short-term predictions. Besides the flows from the previous time step, the flows from time steps 1, 2, 3, 4, and 5 hours before the predicted flow were also used as inputs. This is based partly on the mentioned time of concentration, which is, as was noted, a maximum of 3 hours in this watershed; more time steps (not 3, but 5) were taken, because it is important if there were higher flows in the previous hours, which would indicate that the watershed at that time would be more saturated with water. On the contrary, lower flows signalize a less-saturated watershed, e.g., different initial physical conditions in the watershed.

Table I Statistics of the flows in the gauging stations after dividing the data

FLOWS [$m^3.s^{-1}$]		MAX.	MIN.	AVG.
Q6200-1	TRAIN. (2007-2009)	427.6	1.62	14.07
	TEST. (feb.- dec. 2010)	414.6	3.05	23.26
Q6170-1	TRAIN. (2007-2009)	120.1	0.05	2.96
	TEST. (feb.- dec. 2010)	135.3	0.15	5.38
Q6179-1	TRAIN. (2007-2009)	122.7	0.1	2.99
	TEST. (feb.- dec. 2010)	116.5	0.6	5.37
Q6180-1	TRAIN. (2007-2009)	329.4	0.43	7.14
	TEST. (feb.- dec. 2010)	281.7	0.6	13.31
Q6190-1	TRAIN. (2007-2009)	93.1	0.35	4.38
	TEST. (feb.- dec. 2010)	127.7	0.7	5.84

One additional variable was constructed for each gauging station, which was the difference between the most recent and previous flows at this station. The reason for using this construction is to include some recently observed dynamics of the river flows into the input dataset. On the basis of the considerations just described, the final dataset has 30 features and 35,000 rows. It is a huge data set, and its reduction is desirable, because the computer modelling time could otherwise inappropriately increase.

Eliminating the problem of the "missing values" that were found in the data was the next step in the data preparation phase of the CRISP-DM. The data values could be missing

because they were not measured, are unknown, were corrupted, etc. Data mining methods vary in the way they treat missing values. Typically, they omit any records containing missing values, replace any missing values with the mode or mean, or infer the missing values from the existing values. The imputation of the mentioned missing data in our case is accomplished by linear interpolation, because the character of the flow data allows this opportunity (this is hardly possible, e.g., with precipitation data). Incidentally, this treatment of missing data is also useful in contexts other than in data-driven modelling, which is the theme of this paper, and many other and more sophisticated procedures are available for solving this task.

The next operation in the data preparation should be sampling. As can be seen in Fig. 3, which contains a histogram of the predicted flows, there is a huge amount of relatively low flows, e.g., flows which indicate that nothing important was going on at the time of their measurement in the watershed. In such periods flows neither rise nor fall significantly, so the data in such periods are not very important for flood predictions. This led to the authors' decision to filter out some of them, because of the mentioned computation time problems with large datasets. As can be seen in Fig. 3, high flows are somewhat rare. Because high flows are the most important data in flood predictions, we left all this rare and large flow data in the final training dataset. The filtered rules used for all the flows are in Table II.

Table II Criteria used for the sampling

Intervals of flows [$m^3.s^{-1}$]	Samples in dataset [counts]
$Q < 15$	200
$15 \leq Q < 20$	200
$20 \leq Q < 30$	200
$30 \leq Q < 50$	300
$Q \geq 50$	ALL

Although some features were highly correlated (the correlation coefficients for different features were in a range of 0.208 to 0.996), the authors decided not to remove any features from the data set on this basis, because in the case of Support Vector Machines (SVM) or the ensemble model learning schemes which were applied in this work, any correlation does not have a significant influence on the final degree of precision.

After filtering out some rows from the data as described in the previous paragraph, a training data file reduced in size was obtained (3200 rows).

Some data-driven models could not work with the numeric values of the original data, because the features are not from the same range of values. The normalization of the data is applied in such cases. Normalized data is scaled to fit in a specific range. Each column of the data (the flows from the gauging stations) except for the date are normalized. Each column has its minimum and maximum values, which are defined by the column's range of values. If range [0, 1] is applied, the maximum flow at the gauging station is transformed to 1 and the minimum to the value 0; the other values are normalized by interpolation.

V. Modeling

Hydrological models can be categorized into various categories; one possible classification distinguishes physically-based models, conceptual models and data-driven models. Physically-based models solve exact physically governing equations such as the conservation of mass or momentum equations (differential equations), usually on the basis of spatially distributed inputs. So-called "conceptual" models can be viewed as a combination of mathematical operators, which describe the main features of an idealized hydrological cycle, and it is typical that some of the data they use are not based on real counterparts in nature, e.g., are not directly measurable (usually because of aggregating some processes, etc.). Data-driven models analyse and derive results only from the observed input (e.g., discharges, temperatures, rainfall) and output of a watershed (e.g., a flow); they do not use exact physical descriptions at all.

Two approaches for flood prediction by data-driven modelling are used in this paper: a single model – support vector machines and an ensemble model, which consists of more single models. The results of both approaches are numerical predictions of the flows in the Kysucke Nove Mesto gauging station in cubic metres per second. Both methods are compared with the measured hydrological data of the flows observed.

A. Support vector machines (SVM)

An SVM uses nonlinear mapping to transform the original training data into a higher dimension. The mapping of the original data into this new space is carried out with the help of so-called kernel functions. Within this new dimension, a linear model is constructed. The quality of the estimation of this model is measured by the ε-insensitive loss function, which ignores errors that are situated within a certain distance ε from the true value. As a consequence, an SVM find this model by using not all the input vectors (the data rows), but only the so-called support vectors ("essential" training vectors), which strongly enhance the effectiveness of the algorithm.

Parameter ε controls the width of the ε-insensitive zone, which is used to fit the training data. The value of ε can affect the number of support vectors used to construct the regression function. The larger ε is, the fewer support vectors are selected, which has an influence on a model's degree of precision. On the other hand, a very small ε could lead to over-fitting.

Another useful feature of a SVM is the selection of a model by minimizing structural risk, which corresponds to finding a model which is as simple as possible and is best in terms of any empirical errors in the data. This compromise between empirical and structural minimization leads to the remarkable ability of a SVM to generalize. Parameter C determines the trade-off between the model's complexity (flatness) and the degree to which deviations larger than ε are tolerated in the optimization of the regression function. For example, if C is

too large (infinity), then the objective is to minimize only the empirical risk, without regard to the complexity of the model in the optimization formulation. The mathematical details of this formulation are avoided here. They can be found in the literature, e.g., [5], [10]. The LIBSVM implementation of a SVM is described in [6].

B. Optimization of the SVM's parameters

The proper parameters of a SVM should be searched for in a particular task [11], [12]. The sequence of the practical steps while using the ε-SVR is as follows: selecting a suitable kernel and the appropriate kernel's parameters, specifying the ε parameter, and specifying the capacity C.

The radial basis function was chosen as the kernel function on a trial-and-error basis for which the parameter γ should be specified. A cross-validation methodology with ten folds was used for finding the mentioned parameters of the SVM model. This means that some generator of the possible parameter values is chosen – this could be a grid generator, which produces parameters in a grid manner with predefined steps, or an evolutionary generator, which proposes parameters on the basis of, e.g., genetic algorithms. The latter was used in this study. The task of cross-validation is to identify the best parameters solely on the basis of the training data.

The cross-validation tests the precision of a model on part of the training data. The training data were divided into 10 parts. Nine different parts were used ten times for the training of the ten SVM models, and in every case the unused tenth part was used for the validation (always a different combination of the nine parts plus one). The precision of these ten SVM models was averaged into a single precision coefficient for the given setting of the parameters in the particular iteration in which they were searched for. This precision was expressed by the Nash-Suttcliffe coefficient. The use of cross-validation in the optimization process improves the selection of the parameters in comparison with other methods, e.g., a method based only on a single division of the training data in the training and validation group, as the resulting model is more likely to be objective and stable with regard to the degree of precision for the various data not used in the training process.

Each parameter was searched in the optimization process in a range established on the basis of the authors´ experience and information from the SVM literature.

C. Ensemble models

The authors of the paper wanted to answer the question as to whether improvements in performance are obtained by ensemble modeling of river flow predictions in comparison with each of the ensemble members' performances, in a case where these members are already powerful algorithms with good performances. In the usual ensemble approach many learners (e.g., classification and regression trees) are combined in an attempt to produce a strong learner.

The authors of this article are aware of some degree of subjectivity in the choice of the strong algorithms which were included in the proposed ensemble, but some supporting information in the data mining community exists [7], [12]. A grid search combined with a repeated cross-validation methodology was used for finding the parameters of all the models included in the ensemble. In this approach a set of each model's parameters from a predetermined grid is sent to the parameter-evaluating algorithm. Basically, a 2-times-repeated 5-fold cross validation was used. The data set was divided into 5 subsets, and the training-testing-evaluation was repeated 5 times. Each time, one of the 5 subsets is used as the test set, and the other 4 subsets are put together to form a training set. Then the average error across all 5 trials is computed, and the case with the lowest errors determines the combination of the SVM parameters in an actual repetition. These parameters were used to train the final model. The division of the data into five subsets was repeated differently two times. This repeated k-fold cross-validation is the main reason for the necessity to accomplish sampling of the data (e.g., data reduction) as mentioned in the data preparation section of this paper, because each basic algorithm runs many times in such a strategy. A brief description of the selected algorithms follows.

Generalized linear model with elastic-net (glmnet)

A generalized linear model is fitted in this case with an elastic-net penalty. In a solved flows prediction task the multiple-linear regression is a linear model. The algorithm uses a cyclical coordinate descent in a path-wise fashion, as described in [13]. An elastic-net is a sort of regularization technique, the aim of which is to obtain as simple a model as possible, while keeping its degree of precision on an appropriate level. The application of the regularization leads to models with better generalizations, e.g., predictions on the basis of data which were not used to train the model. An elastic-net penalty function has two roles: controlling the "sparseness" of the solution (the number of coefficients that are non-zero) and controlling the magnitude of the non-zero coefficients ("shrinkage"). In this work the original software provided by the authors of this method was used with its default settings [14].

Gradient boosting machines (GBM)

Gradient boosting machines (GBM) are one of the most powerful and popular boosting methods. A GBM involves fitting a series of trees, with each successive tree being fit to a resampled training set that is weighted according to the accuracy of the previously fitted tree. The original training data is resampled several times, and the combined series of trees forms a single predictive model. GBMs train many models, and each new model gradually minimizes the loss function of the whole system using a gradient descent method, e.g., it builds the model in a stage-wise fashion. The coefficients are fitted incrementally, one at every step. The algorithm may be found in [15]. The GBM model for the analysis of flows reported here was fit using the gbm package in R [16]. A GBM performs shrinkage implicitly: the coefficients are "shrunk" with a Lasso-type penalty with the shrinkage controlled by υ, which was set in this work to 0.01. This fact is understood to be one of the key reasons for the superior performance of the algorithm. The total number of

trees to fit is equal to 700 in this work, and this parameter was found by a grid search. Also, the maximum depth of the variable interactions was found by a grid search with up to 10-way interactions.

Support vector machines (SVM)

The authors developed the ε-SVM model by: 1) selecting a radial basis kernel by a trial-and-error approach and then this kernel's parameter sigma = 0.0005, which was found by a grid search; 2) specifying the ε parameter to be equal to 0.1 (the usually recommended value); and 3) specifying the capacity C = 10.5 by a grid search.

Differential Evolution

The differential evolution (DE) was first proposed by Storn and Price [17] in 1997, as a generic metaheuristic for the optimization of nonlinear and non-differentiable continuous space functions, and it has proven to be very robust and competitive with respect to other evolutionary algorithms. At the heart of its success lies a very simple differential operator, whereby a trial solution vector is generated by mutating a random target vector by some multiple of the difference vector between two other random population members. For the three distinct random indices *i*, *j* and *k*, this has the form

$$y_i = x_i + \hat{f} \times (x_i - x_k) \quad (1)$$

where x_i is the target vector, y_i is the trial vector, and $\hat{f}$ is a constant factor in the range [0, 2] which controls the amplification of differential variations, typically taken as 0.5. If the trial vector has a better objective function value, then it replaces its parent vector. Storn and Price also included a crossover operator between the trial vector and the target vector in order to improve the convergence.

VI. Evaluation

The process of assessing the performance of a hydrologic model involves making some estimates of the "closeness" of the simulated behaviour of the model to the observations (in our case, the stream flow). An objective assessment requires the use of a mathematical estimate of the error between the simulated and observed hydrological variable.

The following evaluation measures have been used to compare the performance of the model, where N is the number of observations, O_i is the actual data, and P_i is the predicted value(s):

The *Nash-Sutcliffe efficiency (NSE)* is a standardised statistical ratio defining the relative magnitude of the variability of the residuals compared with the dispersion of the measured data. The NSE ranges from -∞ up to and including 1, where NSE = 1 means a perfect agreement between the measured and simulated data.

$$E = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2} \quad (2)$$

The *Pearson correlation* coefficient (r) describes the level of collinearity between the simulated and measured data. The correlation coefficient takes values within a range of -1 to 1, and it is a measure of the linear relation between the observed and simulated data. If r = 0, there is no relation. If r = 1 or -1, there is a perfect positive or negative linear relation. Although it is frequently used for the assessment of models, its suitability must be verified (e.g., by a visual inspection of the graphed results), because it does not catch multiplicative or lagged differences between the modelled and measured data. It has the following form:

$$r = \frac{\sum_{i=1}^{N}(O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{N}(O_i - \bar{O})^2}\sqrt{\sum_{i=1}^{N}(P_i - \bar{P})^2}} \quad (3)$$

The last criterion for comparisons is the overall outflow, which in our case represents the overall outflow from the Kysucke Nove Mesto gauging station in the tested season from 1.2.2010 to 31.12.2010. This criterion describes the balance of the model, i.e., the overestimated or the underestimated trend of the overall prediction of the model in comparison with the observed data.

A. Single model - SVM

A single model based on the SVM with optimalised parameters and the reduced training dataset achieved a correlation coefficient with a value of 0.922 and a Nash-Sutcliffe efficiency with a value of 0.894, which confirms a very highly precise solution of this state-of-the-art regression algorithm. The overall outflow underestimated the measured outflow by 48 million m^3. The overall simulated outflow with regard to the SVM model represents a deviation of 7.13% from the measured outflow.

Table III Evaluation of the single model - SVM by r, NSE and overall outflow

	SVM	Measured data
NSE	0.894	
r	0.922	
Overall outflow [m^3]	625.10^6	673.10^6

B. Ensemble model

The ensemble model is proposed to have the following structure:

$$P_{ensemble}^{t} = \sum_{i=1}^{n} \beta_i * P_i^t, \quad (4)$$

where β_i are the weights of the models of which the ensemble consists, and P_i^t are the predictions by these models in time *t*. In this study 7-hour ahead flow predictions by ensemble modeling are evaluated, the flows of which are computed *n* times for each hour, where *n* is the number of models. The weights of every model in the ensemble are proposed to be found by the differential evolution methodology, which was described in the previous section. The utilization of the differential evolutions for this task follows in the next paragraphs.

The problem solved should be defined by the objective function, which is proposed in this paper to have the following form:

$$O_f = 1 - \left(1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2}\right) + |1 - \sum_{i=1}^{n}\beta_i| \quad (4)$$

$$0 \le \beta_i \le 1, \quad (5)$$

where O_i are the observed flows, N is the number of such data, and $\bar{O}$ is their average value. The component in the rounded parentheses is the Nash-Sutcliffe efficiency (NSE), which is used here because it is one of the most commonly used statistics in hydrology. The last component of the objective function (absolute value) forces the sum of the ensemble members weights β_i to be equal to 1. This objective function is proposed to be minimized. In the case of an ideal model, the value of the objective function is zero.

The ensemble modelling involves an evaluation of each of the three single models in the training phase (Table IV). Then the differential evolution algorithm task was to find the best weights of the individual models on the basis of the objective function discussed. The GLMnet was excluded from the ensemble by the zero weight. The ensemble model confirms the assumption of the higher degree of accuracy in the training phase achieved by the combination of more models.

Table IV Evaluation of the computations in the training phase by r and NSE and the final values of the model weights in the ensembles

Model	Weights	NSE	r
GBM	0.3793	0.942	0.971
SVM	0.6207	0.950	0.975
GLMnet	0	0.903	0.950
ENSEMBLE		0.954	0.977

The ensemble model, including the GBM and SVM models in the testing phase, confirmed the results from the training phase, as seen in Table V.

Table V Evaluation of the computations in the testing phase by r and NSE and the values of the model weights in the ensemble

Model	Weights	NSE	r
GBM	0.3793	0.906	0.951
SVM	0.6207	0.894	0.922
GLMnet	0.0000	0.889	0.921
ENSEMBLE		0.911	0.955

The Nash-Sutcliffe efficiency with a value of 0.911 and a correlation coefficient with a value of 0.955 confirms the ensemble model as the best solution. The predicted overall outflow of 663 million m^3 in comparison with the measured outflow of 673 million m^3 is underestimated by 10 million m^3. The overall simulated outflow with regard to the SVM model represents a deviation of 1.49 % from the measured outflow.

VII. Conclusion

This paper generally describes the possibilities for and suitability of the application of data mining methods in hydrology. In the application part it deals with predicting flows in the Kysuca River basin in Slovakia. The CRISP-DM process is described and applied in the first part. Papers on data mining often focus on the statistical and machine learning algorithms used to make predictions, classifications, etc. Real-world data miners, however, spend most of their time preparing and cleaning the data. That is the reason why the authors decided to devote more space to these parts of the data mining process in this paper and have emphasized the importance of these procedures.

Two types of data-driven modelling are compared, i.e., a single data-driven model and an ensemble model. To make the comparisons fair, a state-of-the-art support vector machine model was selected as the single model. In accordance with the expectations, the flows predicted by this model for 7 hours ahead were computed with a very good degree of accuracy. But in the following experiment, when the SVM model was used as a part of the ensemble model together with other single models, an even higher degree of accuracy was achieved than when the single algorithm was used. On this basis the application of the proposed ensemble methodology could be recommended as a promising alternative for flow predictions in flood warning systems.

According to the so-called "no free lunch" theorem, it is never clear in advance which machine learning algorithm suits best for a particular task. For this reason it is usually necessary to try more algorithms. Instead of selecting and using only the best algorithm, it is better to compose an ensemble prediction, which, as has been shown in this paper, is relatively easy to accomplish when tuned algorithms are already available.

Acknowledgment

This work was supported by the Slovak Research and Development Agency under Contract No. APVV-0496-10 and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, Grant No. 1/1044/11 and 1/0908/11.

References

[1] C. R. Chibanga, J. Berlamont, J. Vandewalle, "Use of Neural Networks to Forecast Time Series: River Flow Modeling," *3rd WSEAS International Conference on Neural Networks and Applications*, Interlaken, Switzerland, February 11-15, 2002, ISBN: 960-8052-48-3

[2] D. M. Oprea, A. Matei, "Applying Artificial Neural Networks in Environmental Prediction Systems," *Proceedings of the 11th WSEAS International Conference on Automation & Information*, 2010, pp. 110-115, ISBN: 978-960-474-193-9

[3] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining," *Journal of Data Warehousing*, Vol. 5, No. 4, 2000, pp. 13-22

[4] V. P. Singh, D. A. Woolhiser, "Mathematical Modeling of Watershed Hydrology," Journal of Hydrological Engineering, Vol. 7 No. 4, 2002, pp. 269–343.

[5] J. Shawe-Taylor, N. Cristianini, "*Support Vector Machines and other kernel-based learning methods,*" Cambridge University Press, 2000

[6] C. C. Chang, C. J. Lin, "LIBSVM: A Library for Support Vector Machines," Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001

[7] C. Rich, A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," *Proceedings of the 23rd international conference on Machine Learning.* ACM, 2006, pp. 161-168

[8] A.P. Kraipeerapun, S. Amornsamankul, "Prediction of WQI for Tha Chin River Using an Ensemble of Support Vector Regression and Complementary Neural Networks," *Recent Advances in Information Science, Proceedings of the 7th European Computing Conference* (ECC '13), Dubrovnik, Croatia, June 25-27, 2013, ISBN: 978-960-474-304-9, eds. D.Boras et al.

[9] B. C. Hernández-Espinosa, J. Torres-Sospedra, M. Fernández-Redondo, "Combination Methods for Ensembles of RBF Networks," *Proceedings of the 6th WSEAS Int. Conf. on NEURAL NETWORKS*, Lisbon, Portugal, June 16-18, 2005, pp. 140-145, ISBN: 960-8457-24-6

[10] V. Vapnik, "*Statistical learning theory*." 1998

[11] H. B. Alwan, K. R. Ku-Mahamud, “Incremental Continuous Ant Colony Optimization Technique for Support Vector Machine Model Selection Problem,” Mathe*matical Methods for Information Science and Economics*, 2012, pp.165-170, ISBN: 978-1-61804-148-7

[12] C. Rich, N. Karampatziakis, A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," *Proceedings of the 25th international conference on Machine Learning*. ACM, 2008, pp. 96-103

[13] J. Friedman, T. Hastie, R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software* 33.1, 2010

[14] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, “Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent,” *Journal of Statistical Software*, 39(5), 2011, pp.1-13

[15] T. Hastie, et al., "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer* 27.2, 2005, pp. 83-85

[16] G.Ridgeway, “Generalized Boosted Models: A guide to the gbm package,” Update1.1, 2007

[17] R. Storn, K. Price, “Differential evolution: A simple and efficient heuristic for global optimization over continuous spaces,” *Journal of Global Optimization*, 11, 1997, pp. 341-359