Environmental & Socio-economic Studies

DOI: 10.1515/environ-2016-0010

Environ. Socio.-econ. Stud., **2016, 4, 2: 34-42**

Original article

Machine-learning methods in the classification of water bodies

Marek Sołtysiak*[1], Marcin Blachnik[2], Dominika Dąbrowska*[1]

[1] *Department of Hydrogeology and Engineering Geology, Faculty of Earth Sciences, University of Silesia, Będzińska Str. 60, Sosnowiec, Poland*

[2] *Department of Industrial Informatics, Silesian University of Technology, 40-019 Katowice, Krasińskiego Str. 8, Poland*

*E–mail address (*corresponding authors): marek.soltysiak@us.edu.pl, ddabrowska@us.edu.pl*

ABSTRACT

Amphibian species have been considered as useful ecological indicators. They are used as indicators of environmental contamination, ecosystem health and habitat quality., Amphibian species are sensitive to changes in the aquatic environment and therefore, may form the basis for the classification of water bodies. Water bodies in which there are a large number of amphibian species are especially valuable even if they are located in urban areas. The automation of the classification process allows for a faster evaluation of the presence of amphibian species in the water bodies. Three machine-learning methods (artificial neural networks, decision trees and the k-nearest neighbours algorithm) have been used to classify water bodies in Chorzów – one of 19 cities in the Upper Silesia Agglomeration. In this case, classification is a supervised data mining method consisting of several stages such as building the model, the testing phase and the prediction. Seven natural and anthropogenic features of water bodies (e.g. the type of water body, aquatic plants, the purpose of the water body (destination), position of the water body in relation to any possible buildings, condition of the water body, the degree of littering, the shore type and fishing activities) have been taken into account in the classification. The data set used in this study involved information about 71 different water bodies and 9 amphibian species living in them. The results showed that the best average classification accuracy was obtained with the multilayer perceptron neural network.

KEY WORDS: water body, k-nearest neighbour algorithm, artificial neural network, decision tree, amphibians

ARTICLE HISTORY: received 9 March 2016; received in revised form 7 April 2016; accepted 27 May 2016

1. Introduction

Water bodies constitute a valuable ecosystem with favorable conditions for preserving biodiversity (SCHEFFER & VAN NESS, 2007). Within their surroundings specific natural and anthropogenic features occur. Amphibians, as one of the most vulnerable animal phyla, and protected by law (IUCN, 2010), are sensitive to transformation of the environment. Simultaneously, they are also a very sensitive indicator of the quality of the aquatic environment. Small water bodies (below 1 ha) are particularly important for amphibians as their breeding sites. Three amphibian species (marsh frog, edible frog, grass frog) also use water bodies as a wintering place (SOŁTYSIAK & DĄBROWSKA, 2014).

The development of agriculture as well as urbanization has caused the transformation of aquatic ecosystems in urban areas. The status of water bodies in Europe is far from optimal. There is a disturbing trend towards decreasing numbers of water reservoirs. For example on the greater part of European countries liquidated 40-90% of small water bodies (OERTLI ET AL., 2002).

In Poland, the Upper Silesia Agglomeration has experienced the biggest changes in aquatic ecosystems (MACHOWSKI & NOCULAK, 2014). A high concentration of several branches of industry and a great number of coal mines have caused the formation of subsidence troughs and the degradation of water quality and the elimination of other water bodies. According to the Central Statistical Office in Poland (ROCZNIK STSTYSTYCZNY, 2015), the Upper Silesia Agglomeration which is one of the most industrialized areas in Poland, consists of 19 cities, covering an area of 1.471 km^2 (Fig. 1).

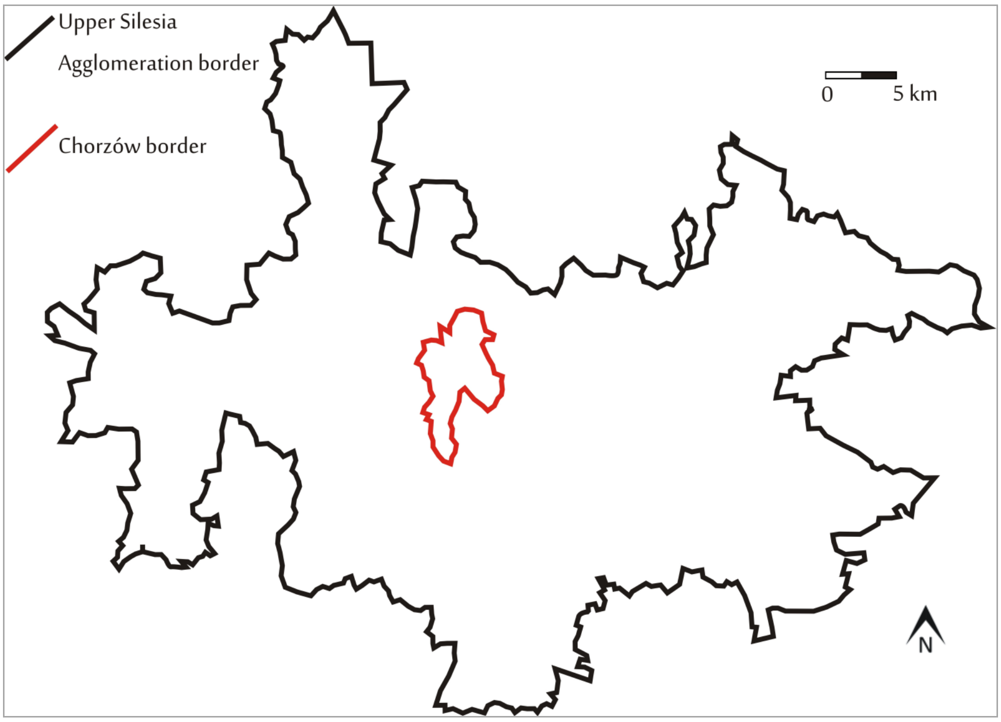

Fig. 1. The location of the study area

There are 4773 water bodies in the Upper Silesia Agglomeration therefore this region was named Upper Silesian Anthropogenic Lakeland (Rzętała, 2008). Hydrogenesis of water bodies is varied but they mainly have anthropogenic characteristics.

The type of anthropogenic water reservoirs is influenced by a few factors such as: the type of littoral margin, the degree of vegetation, distance from the water body to urban areas or roads and can also be influenced by the number of various ecological indicators. It is reasonable to perform such classification research for water bodies, but because of the lack of time and costs, such studies are not carried out. It is therefore possible to overlook the real impact of urbanization on the populations of amphibians, which may result in fragmentation of natural habitats and, as a consequence, to a decrease in their population. For this reason, it would be very important to assess the significance of potential amphibian sites by using modern computational methods and techniques that can operate with less empirical data.

2. Theoretical background

As described above, the automation of water body classification is desired from many perspectives, and the most appropriate tool which could be applied for that purpose are machine learning algorithms, especially supervised machine learning (SML). SML consists of two main stages: building a prediction model and applying that prediction model to classify new examples. More formally building the prediction model is responsible for finding a mapping function $y = f(\mathbf{x})$ which maps an input example $\mathbf{x}$ to a value y. Basically there are two types of SML: classification and regression. The first one finds a mapping when $y \in \{C_1, C_2, \ldots, C_k\}$ where C denotes one of k symbols, and the second when $y \in R$.

To find that mapping function a training data set is needed which consists of n pairs $\mathbf{T} = \{\langle \mathbf{x}_1, y_1 \rangle, \langle \mathbf{x}_2, y_2 \rangle, \ldots, \langle \mathbf{x}_n, y_n \rangle\}$. That allows the training algorithm to tune the internal parameters of the prediction model (also called decision model). When the prediction model is trained, so the function $f()$ is determined, then comes the second stage where this function is applied to classify new examples $\mathbf{x}$, so to find associated label value y. It is important to note that $\mathbf{x}$ is usually represented as a fixed size vector which consists of m variables also called features or attributes. The process of using a machine learning algorithm is presented in Fig. 2.

There are two indicators which influence the quality and accuracy of the final decision model. These are: the type of model, and the quality of the training dataset which depends on the size of the training set n, quality and size of the feature space m. The first indicator must be adjusted empirically as it is almost impossible to apriori identify which type of model would be suitable for a given dataset. There are many types of models such as feed forward neural networks, decision trees, distance based methods such as kNN method or kernel methods such as Support Vector Machine. The second indicator is even more important as even the best model requires a good quality of training data (ZHU AND XIANGXIN, 2012).

In our experiments described in section 5 (Experiments and results) we tried to build and assess three different types of SML models, namely: MLP neural network trained with a momentum algorithm, C4.5 decision tree and nearest neighbor model. The dataset was created based on data collected in our research related to the empirical assessment of Chorzów water bodies conducted in the year 2004. The feature space was determined by our own experience and properties which influence the occurrence of different species of amphibians.

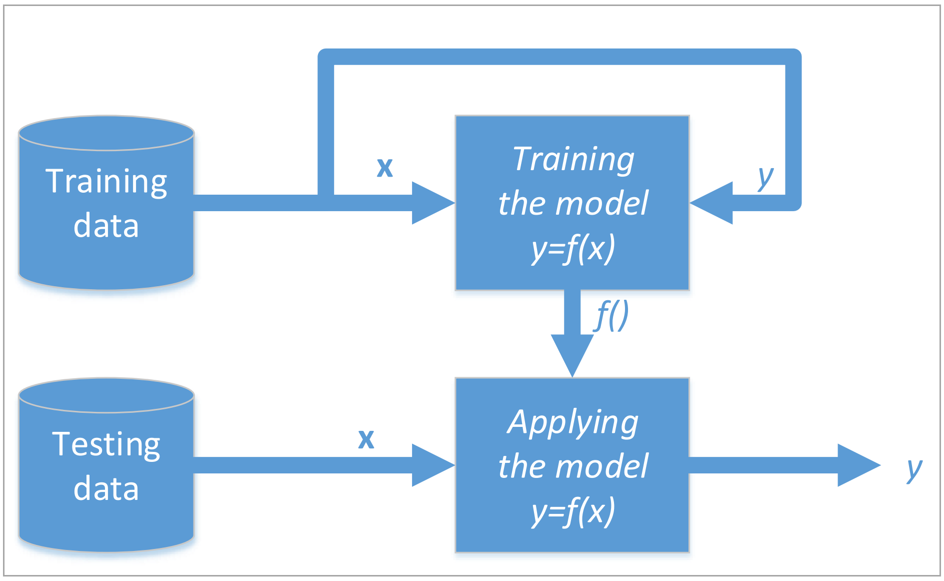

Fig. 2. Process of supervised machine learning

3. Study area

The study site – Chorzów city is located in the southern part of Poland. The area of Chorzów is equal to 33.5 km^2 with a population of 110 000 people (ROCZNIK STATYSTYCZNY POLSKI, 2015). The characteristic feature of the spatial structure is the presence of a closely built-up city centre and peripheral areas with less development (Fig. 3). There are three forested areas in Chorzów: the Żabie Doły (Frog Pits) in the north, the Silesian Park in the central-eastern and the forest complex in the south (SOŁTYSIAK & DĄBROWSKA, 2015). Table 1 contains particular information about the spatial structure of the city.

Table 1. Spatial structure of the city Chorzów (Aglomeracja śląska w liczbach, 2006)

Type of land use	Area [km^2]
Built-up areas (industrial areas included)	15.5
Agricultural land	2.64
Forest land	2.37
Urban green areas	6.12
Wastelands	2.86
Allotments	1.93
Heaps	0.50
Recultivated land	0.32
Land under water	0.70
Other areas (with transport areas)	3.60

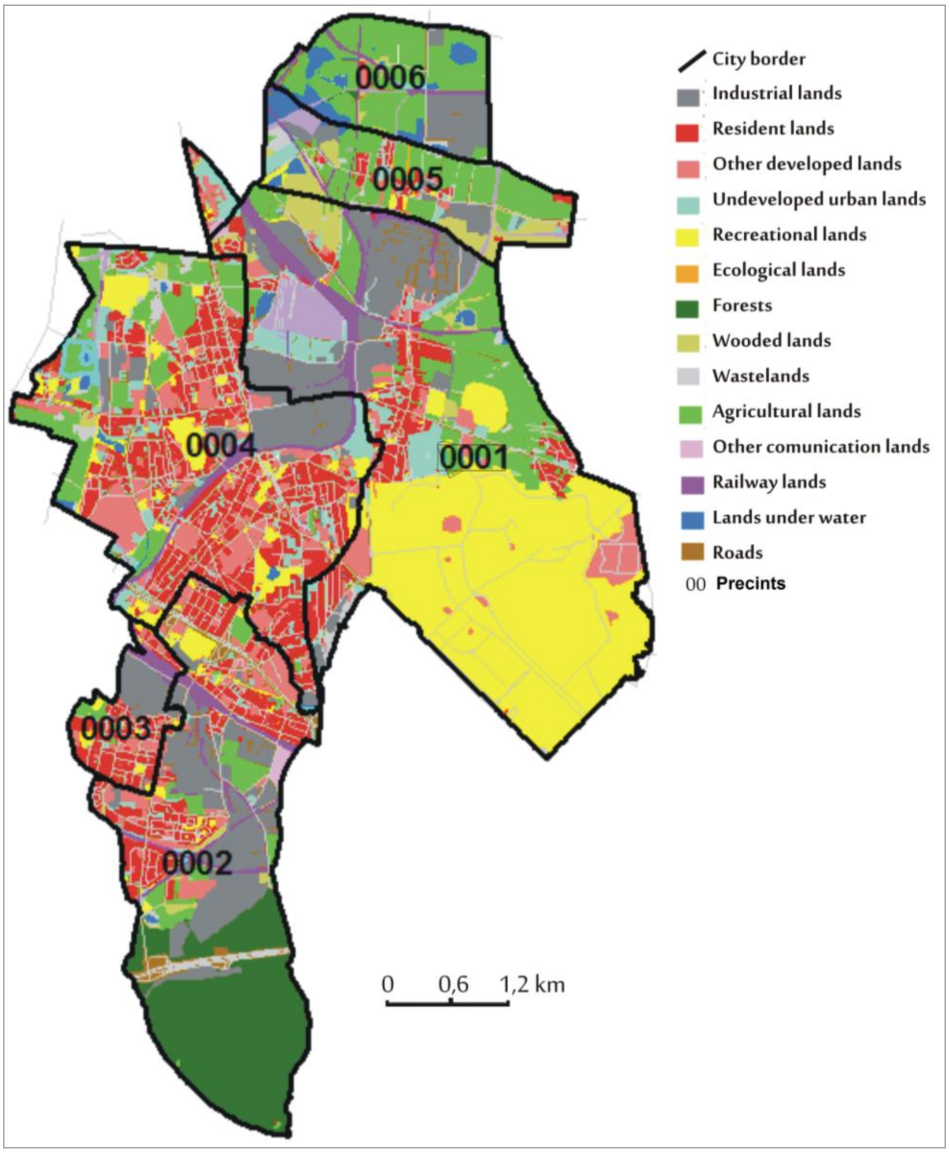

Fig. 3. Spatial structure of Chorzów (e-odgik.chorzow.eu)

4. Material and methods

A review of available topographic maps of 1:10,000 (Mapa topograficzna, 1993) scale and orthophotos (www.geoportal.gov.pl) and own field research gave a indication of the actual number of water bodies in the study area. The water bodies were described using 7 features and a few categories for each feature. These are:
Feature 1: the type of water body (1 - a subsidence or post-exploited water body, 2 - a pond, 3 - a tank, 4 - water body located near a house, 5 – water body in a garden, 6 - technological water body, 7 - water body made from concrete, 8 - water body in a green urban area),
Feature 2: the degree of vegetation (from 0 - without to 4- with a lot),
Feature 3: the type of surroundings (1 - forest, 2 - wastelands or meadows, 3 - allotments, 4 - parks, 5 - urban areas, 6 - roads, 7 - agricultural land),
Feature 4: the water use (1 -waste, 2- recreation, 3 - usable, 4 - industrial),
Feature 5: the degree of littering (0 - without, 1 – with a lot of),
Feature 6: the type of littoral zone (1 - natural, 2 - concrete),
Feature 7: angling occurrence (from 0 - without to 3 – with a lot of).

The class label value describes the presence of nine amphibian species (Green frog, Grass frog, Common toad, Green toad, Spade foot, Ordinary newt, Great crested newt, Tree frog, Fire-bellied toad). As different species can occur together this leads to a multi-label learning problem and in our case we defined nine binary label attributes each for given species of amphibians. The herpetological study was based on the results of own research (Sołtysiak, 2004) and covered the whole area of the city.

Three supervised machine learning algorithms were used in the calculations. We have chosen C4.5 as an algorithm for building the decision tree, k-nearest neighbor algorithm and Multi Layer Perceptron artificial neural network.

A decision tree internally represents induced knowledge in a form of a tree structure where the internal nodes of the tree represent a rule condition, and leaf nodes represent final decision. Internal nodes (decision nodes) represent a test on an attribute (usually attribute and threshold or attribute and *contains* a relation depending on feature type); and the leaf nodes represent a class label (DAI & JI, 2014). A classification rule is saved in each path from the root node to leaf node.

The raining data set is applied to search for the best splitting attribute for each node. The algorithm recursively starts from the root node, and then, the input dataset is partitioned into a few subsets (depending on the type of the tree – usually into two parts), which are delivered to the new nodes, and the procedure repeats. The algorithm completes if all training instances which fall into a node belong to one class (QUINLAN, 1993; DEVEZE & FOUQUIN, 2005).

K- nearest neighbor is an algorithm that stores all possible training examples and classifies new examples based on a similarity to the examples stored during training (LOPEZ ET AL., 2001). An example is classified by the majority of its neighbors, and the label is determined by the most common class among its k nearest neighbors.

The nearest neighbors of the query instance are determined by some distance function. For continuous attributes usually Euclidean distance is used, but also other distance or similarity measures are possible such as: Manhatan, Minkowski, Canberra, etc. This simple algorithm is classified as one of 10 best known algorithms in data mining (WU ET AL., 2008), but it requires the determination of the appropriate value of considered nearest neighbours (*k*). Artificial neural networks are an information processing technique inspired by the nervous systems. One of the most popular types of the artificial neural network is multilayer perceptron (Fig. 4) (RUCK ET AL., 1989).

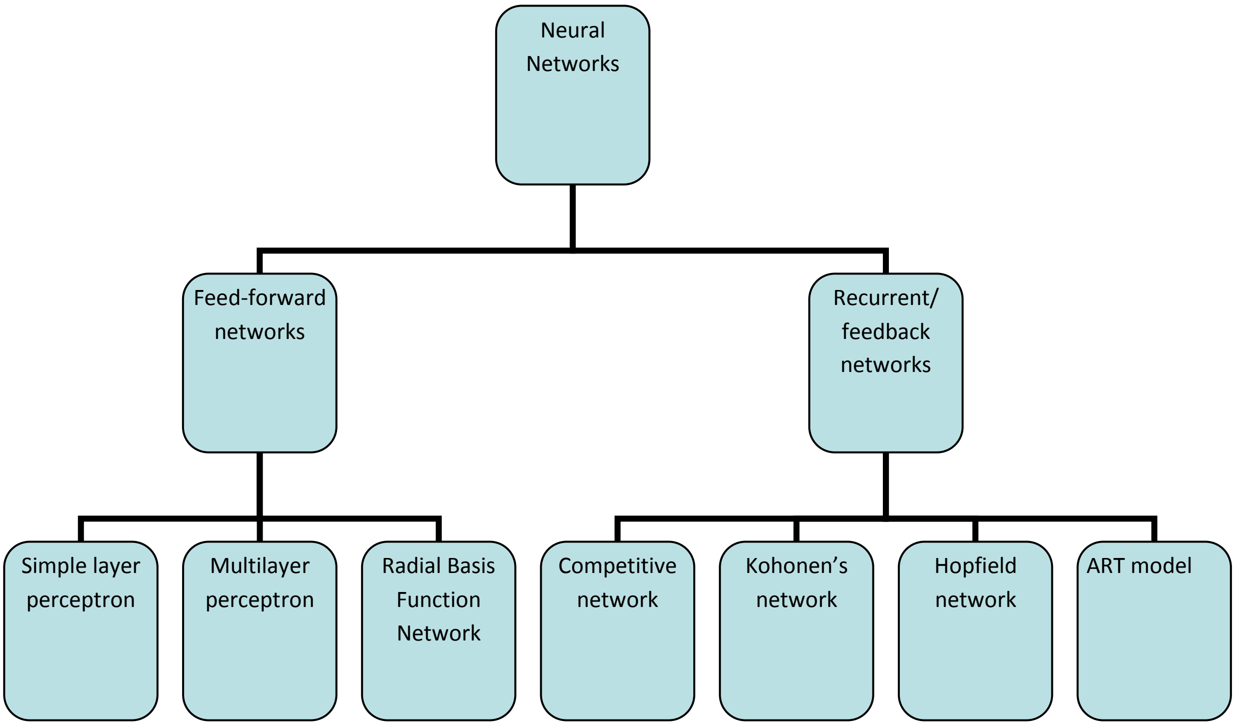

Fig. 4. The architecture of artificial neural networks (after Jain et al., 1996)

The multilayer perceptron consists of a set of simple nodes, so called neurons organized in layers. A single neuron combines a linear weighted sum of input activations with a nonlinear operator (usually sigmoidal function) (GARDNER & DORLING, 1998), and training of a single neuron is based on adjusting the weight factors of the sum. A multilayer perceptron consists of several layers of neurons – an input layer which serves to pass the input data, one or more hidden layers and an output layer. This network learns adjusting neuron weights starting from the output neurons, and back propagating the value of error to training the neurons in the hidden layers using so called back-propagation algorithm (RUMELHART, 1987).

5. Experiments and results

C4.5 algorithm, k-nearest neighbours algorithm and multilayer perceptron networks were used to classify 71 water bodies in Chorzów on the basis of the diversity of amphibian species occurring in them. All calculations were performed using RapidMiner 7.0, an environment for predictive analytics (Blachnik & Kordos, 2015). The data set for the water bodies in Chorzów is presented in Appendix 1. The data mining process used in our experiments is presented in Fig. 5.

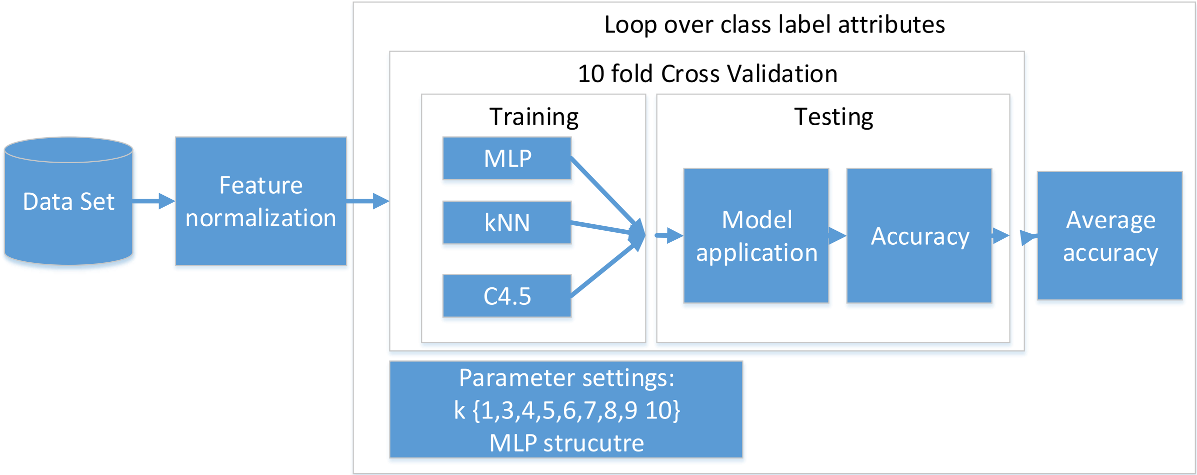

Fig. 5. Structure of the data mining process

The process starts by loading the data, then each attribute is normalized into the range [0,1], next the loop starts which iterates over different class label attributes, such that in our case for each species of amphibian an independent prediction model was built and evaluated. The performance and quality of the prediction model was assessed using a 10-fold cross-validation test. This procedure repeats 10 times the process of building the prediction model and evaluating its performance each time on an independent test set sampled from the input data. Finally after the 10-fold cross validation we obtain 10 different values of the performance measure. In our calculation we used *classification accuracy* defined as a ratio between correctly classified instances and all examples used in the test set. Finally these values are averaged and returned as the output of the calculations. To find the best model the cross-validation test was repeated for each parameter setting such as *k* in kNN or different numbers of neurons, and layers of neurons in the case of MLP network. The best results obtained for each algorithm are presented in Table 2. The best results obtained for each species of amphibian are marked in bold. This table also includes an extra row which represents a base-rate which shows the results of the majority classifier. These results are used as an indicator to check how much the algorithm was able to learn.

Table 2. The accuracy average results for each method

Green frog	Grass frog	Common toad	Green toad	Spade foot	Ordinary newt	Great crested newt	Tree frog	Fire-bellied toad	Algorithm
0.6714	**0.6714**	**0.7429**	0.7000	**0.8571**	**0.5857**	**0.7714**	**0.9143**	**0.9714**	MLP
0.7571	0.6143	0.5571	0.6857	0.8000	0.4714	0.6429	**0.9143**	**0.9714**	C4.5
0.6140	**0.6714**	0.7000	**0.7430**	0.8430	0.5140	0.7570	**0.9143**	**0.9714**	kNN
0.6000	0.5290	0.5860	0.6140	0.7290	0.5140	0.7570	**0.9143**	**0.9714**	Base Rate

Analysis of the results provided in Appendix allows us to conclude that in almost all of the cases the MLP neural network performed best. Except green frog and green toad it significantly outperformed all other methods. The worst results were obtained for the C4.5 decision tree, which often achieved an accuracy below the base-rate such as in the case of common toad or ordinary and great crested newt. The kNN classifier performed in-between these two classifiers. In one case it outperformed all competitors and in another case it achieved identical results to the MLP neural network. For two species, namely tree frog and fire-bellied toad all methods achieved an accuracy equal to the base rate, what disqualifies all obtained models, as it couldn't induce any important knowledge. The source of this defeat is in the data, because both of these species appear only in 7 and 2 out of the 71 water bodies respectively, that is in 9.8% and 2.8% of the cases in the dataset which is very rare.

6. Summary and conclusions

Water bodies are characteristic features of the Chorzów area. Water bodies can also be the habitat for amphibians in the city environment. The classification of water bodies can be conducted using supervised machine learning methods. The results obtained show that the process of automatic classification is possible and reasonably accurate. It must be noted that the dataset used in our experiments only included 71 examples, so extending the dataset should lead to further improvement of the results. For example we can observe this phenomenon on results obtained on tree frog and fire-bellied toad, for which it was impossible to build an accurate model as they appeared very rarely. Generalizing the obtained results we can say that the MLP network achieved the best results, and can be treated as a first choice method for further research.

The next step of our research is to collect more data extending the investigation area to other cities and non-urbanized areas of Upper Silesia. This should allow us to build the learning curve (Perlich, 2011) which may indicate how the classification accuracy depends on the number of training data, and further improve the classification quality.

Appendix 1

Type	Vegetation	Surroundings	Use	Littering		Angling	1	2	3	4	5	6	7	8	9*
1	3	3	1	1	1	1	1	1	1	1	1	1			
1	2	5	1	1	1	1	1		1						
8	1	3	2	0	2	1	1	1	1	1					
8	2	3.5	1	0	1	0	1	1			1	1	1		
1	1	3.5	1	1	1	2	1	1	1			1			
1	1	3	2	0	1	3	1		1						
1	2	3	1	0	1	1	1	1			1		1		
5	2	3	2	0	1	1	1			1		1	1		
2	2	2	1	1	1	0	1	1							
1	2	3.6.2	1	1	1	1	1	1	1		1				
8	2	4.6	1	1	1	0	1	1	1			1	1		
1	2	2.3	1	1	1	0	1	1		1	1	1	1		
5	2	3	3	1	1	0				1					
1	2	6.5	1	1	1	1	1	1	1	1					
1	1	2	2	0	1	3		1	1						
1	4	2.8	1	0	1	0	1	1		1	1	1			
1	2	2.8	1	1	1	2	1			1	1				
1	1	2	1	0	1	1				1		1			
1	1	2	1	1	1	1				1	1	1			
1	1	2	1	1	1	1				1	1				
1	2	2	1	1	1	2	1				1	1			
1	1	2	1	1	1	2									
1	2	2	1	1	1	2	1								
1	1	2	2	0	1	3									
1	2	2	1	0	1	1	1								
1	4	2.8	1	0	1	0					1	1			
1	3	2.8	1	0	1	0	1			1				1	
1	3	8	1	0	1	0	1	1		1	1	1			1

5	0	5.6.2	4	0	2	0				1		1			
1	3	2.8	1	0	1	0	1	1		1	1	1			
1	1	3.2.5	1	0	1	0	1	1	1	1	1	1			1
1	4	3	1	1	1	0		1	1	1					
1	1	3	2	0	1	2	1	1	1	1					
1	3	2.8	1	0	1	0	1	1		1	1	1			
1	3	2.8	1	0	1	0	1	1	1		1				
1	3	2.8	1	0	1	0					1				
1	4	2.8	1	0	1	0	1	1		1	1				
1	1	1.5.6	1	0	1	0		1				1	1		
1	2	4	2.1	0	1	3		1	1				1		
1	1	4	1	0	1	0		1				1	1		
1	1	4	1	0	1	0						1			
1	1	4	1	0	1	0						1			
1	2	4	1.2	0	1	1	1	1	1						
1	4	4	1.2	0	1	0	1	1	1			1	1		
1	2	4	1	0	1	0	1	1				1	1		
6	1	4	2	0	2	0	1	1	1			1	1		
1	1	4	2	0	1	0	1	1	1						
1	2	4	1	0	1	0	1	1	1			1	1		
1	1	4	2	0	1	1									
1	2	4	2	0	1	0	1	1	1						
7	1.5	4	2	0	2	0				1		1			
7	1.5	4	2	0	2	0				1					
1	1	4	2	0	1	0			1						
1	2	2.4	1	0	1	0	1	1	1			1	1		
1	1	4	2	0	1	0		1	1	1		1			
7	0	4	2	0	2	0			1	1	1	1	1		
1	1	4	2	0	1	0			1						
1	1	4	2	0	1	1	1			1		1			
6	0	5.4	4	0	2	0	1			1					
6	0	5	4	0	2	0				1				1	
8	2	3.2	1	3	1	0	1		1			1		1	
1	1	3	2	0	1	0									
2	2	2	1	0	1	0						1	1		
3	3	1	4	0	1	0	1	1	1						
1	3	1	1	0	1	0						1	1		
1	1	2	1	2	1	0	1	1	1			1		1	
1	2	2	1	0	1	0	1	1	1					1	
6	2	1	4	0	1	0	1					1	1		
1	2	1	1	0	1	0		1							
1	2	1	1	0	1	0	1	1						1	
1	2	1	1	0	1	0	1	1				1		1	

*Labels: 1 - Green Frog, 2 - Grass Frog, 3 - Common Toad, 4 - Green Toad, 5 - Fire- bellied toad, 6 - Ordinary Newt, 7 - Great crested newt, 8 - Tree frog, 9 - Spade foot

References

Aglomeracja śląska w liczbach. Główny Urząd Statystyczny, Katowice, 2006.

Blachnik M., Kordos M. 2015. Information Selection and Data Compression Rapid Miner Library. [in:] *Machine Intelligence and Big Data in Industry*. Springer, 2016.

Bruyne S. 2010. *Process, Data and Classifier Models for Accessible Supervised Classification Problem Solving*. VubPress.

Cover T., Hart P. 1967. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory*, 13,.1: 21-27.

Dai W., Ji W. 2014. A Map Reduce Implementation of C4.5 Decision Tree Algorithm. *Int. J. Database Theory and Appl*., 7, 1: 49-60.

Deveze B., Fouquin M. 2005. Datamining C4.5 – DBSCON, PROMOTION, SCIA Ecole pour. I' informatique et techniques avancees.

Gardner M., Dorling S. 1998. Artificial neural networks (the multilayer perceptron) – a review of applications in the atmospheric sciences. *Atmos. Environ*., 32, 14/15: 2627-2636.

Gong P., Howarth P.J. 1990. An assessment of some factors influencing multispectral land-cover classification. *Photogram. Eng. Remot. Sens.,* 56: 597-603.

Han E. 1999. Text categorization using weight adjusted k-nearest neighbors classification. PhD thesis, Univ. Minnesota.

Hodge V., Austin J. 2004. A Survey of Outlier Detection Methodologies. *Artif. Intel. Rev*., 22, 2: 85-126.

http://e-odgik.chorzow.eu/GeoOsrodek/mapa.aspx?wybor=no&control=no&idProject=443

IUCN Red List - Search Results". IUCN Red List of Threatened Species. Version 2010.3. IUCN. Retrieved September 8, 2010.

Jain A., Mao J., Mohiuddin K. 1996. Artificial neural networks – a tutorial. *Computer*: 31-44.

Kotsiantis S.B. 2007. Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 31: 249-268.

Lopez H., Ek A., Bauer M. 2001. Estimation and mapping of forest stand density, volume and cover type using k – nearest neighbors method. *Remot. Sens. Environ*., 77, 3: 251- 274.

Machowski R., Noculak M. 2014. Anthropogenic change in water bodies in the southern part of the Silesian Upland. *Limnological Rev*., 14, 2: 93-100.

Mapa topograficzna 1:10 000, arkusze: Świętochłowice, Chorzów, Ruda Śląska-Kochłowice, Chorzów Batory. CODGIK, 1993.

Nourani V., Alami MT., Aminfar MH. 2008. A combined neural-wavelet model for prediction of watershed precipitation. Ligvanchai, Iran. *J. Environ. Hydrol*., 16: 1-12.

Nourani V., Komasi M., Mano A. 2009. A Multivariate ANN-Wavelet Approach for Rainfall- Runoff Modeling. *Water Resour. Manage*., 23: 2877-2894.

Oertli B., Joye D.A., Castella E., Juge R., Cambin D., Lachavanne J.B. 2002. Does size matter? The relationship between pond area and biodiversity. *Biol. Conserv*., 104: 59-70.

Perlich C. 2011. Learning Curves in Machine Learning. [in:] Sammut C., Webb G. I. (eds.) *Encyclopedia of Machine Learning*. Springer: 577-580.

Quinlan J.R. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann.

Rocznik Statystyczny Polski. Główny Urząd Statystyczny, Warszawa, 2015.

Ruck D. W., Rogers S. K., Kabrisky M. 1989. Tactical target recognition: Conventional and neutral network approaches. *Proc. 5th Ann. Aerospace Appl. Artif. Intel. Configure*: 247-253.

Rumelhart D., McClelland J., PDP Research Group. 1987. Parallel Distributed Processing: Explorations in the Microstructure of Cognition.

Rzętała M. 2008. *Funkcjonowanie zbiorników wodnych oraz przebieg procesów limnicznych w warunkach zróżnicowanej antropopresji na przykładzie region górnośląskiego*. Uniw. Śląski, Katowice.

Sammut C., Webb G. (eds.) 2010. *Encyclopedia of Machine Learning*. Springer, p. 1031.

Scheffer M., van Ness E. 2007. Shallow lakes theory revisited: various alternative regimes driven by climate, nutrients, depth and lake size. *Hydrobiol. Bul*., 584.

Sołtysiak M. 2004. Inwentaryzacja miejsc występowania płazów w Chorzowie – wstępne wyniki badań. [in:] *Biologia płazów i gadów – ochrona herpetofauny*. Uniw. Pedagogiczny w Krakowie, Kraków: 122-126.

Sołtysiak M., Dąbrowska D. 2014. Is there a space for amphibians in the space of Chorzów city and Sosnowiec city? *Urban Fauna,* Univ of Sci. and Techn., in Bydgoszcz, Bydgoszcz: 161-168.

Sołtysiak M., Dąbrowska D. 2015. Presence of small reservoirs in the city space on the example of Chorzów. *PhD Interdiscipl. J*., 1, Gdańsk: 27-36.

Wu X., Kumar V., Quinlan J., Ghos J., Yang Q., Motoda H., McLachlan G., Ng A., Liu B., Yu P., Zhou Z., Steinbach M., Hand D., Steinberg D. 2008. Top 10 algorithms in data mining. *Knowl. Inf. Syst*., 14: 1-37.

www.geoportal.gov.pl

Zhang S., Zhang C., Yang Q. 2002. Data Preparation for Data Mining. *Appl. Artif. Intel*., 17: 375-381.

Zhu, Xiangxin, et al. 2012, "Do We Need More Training Data or Better Models for Object Detection?." *BMVC*. Vol. 3.