International Scholarly Research Network
ISRN Soil Science
Volume 2012, Article ID 610567, 9 pages
doi:10.5402/2012/610567

Research Article

An Ensemble of Neural Classifiers and Constructivist Algorithms in the Identification of Agricultural Suitability Complexes of Soils on the Basis of Physiographic Information

Stanislaw Gruszczynski

Faculty of Mining Surveying and Environmental Engineering, AGH University of Science and Technology, Al. Mickiewicza 30, 30-059 Cracow, Poland

Correspondence should be addressed to Stanislaw Gruszczynski, sgrusz@agh.edu.pl

Received 3 January 2012; Accepted 19 February 2012

Academic Editors: M. Cox, W. Ding, and Z. He

The ensemble of classifiers for identification of agricultural suitability of soils on the basis of physiographic information was created in accordance with the *stacking* algorithm. It is comprised of five neural networks of various structures. The deciding element was a neural classifier optimised on the basis of input vectors composed of the indications of five classifiers making up the lower level. Among the architectures studied, the best result was achieved using the Radial Basis Function network as the decisive classifier, composed with the use of the constructivist Feature Space Mapping algorithm. In this configuration, the group correctly identified more than 99% of the elements of the validation set. The models may be used as tools for predicting expected soil condition, which is helpful in assessment of the range of substantial transformations.

1. Introduction

Classification algorithms are an essential component of spatial information systems as well as of packages for remote sensing data processing. They are designed to identify classes of various objects that can be distinguished based on the information stored in spatial databases. An object class can be, for example, the way in which land is used (arable area, greens, forest, water, developed area, etc.), its habitat type, soil type, and other similar environmental components that can be designated with an unambiguous label, placing them on a specified nominal scale.

The objective in seeking new models and variants is to increase the reliability of classification and identification. Among other uses, neural networks have been employed for several decades for the modelling of phenomena and processes [1–3]. These are evolutionary algorithms classified as artificial intelligence, computational intelligence, and data mining algorithms. The development of architectures and methods for optimisation and neural network assessment, including their use as classifiers, shows both advantages and drawbacks to their applications [4, 5]. Probably the most obvious drawback of evolutionary algorithms is the difficulty in extracting the formal rules underlying the correct classification made by the decision model: this property is referred to as the "black box." Another disadvantage is the phenomenon of specialisation, involving better suitability of some architectures and poorer suitability of others for solving specific tasks [3], in which this usefulness cannot be determined in advance. It is usually impossible to predict which of the potentially usable models is going to perform well in a particular task. Traditionally, many attempts are made to construct models differing in the structure, size, and parameters of the optimisation algorithm with the aim of achieving satisfactory results. The observation that in many problems, some models that fare poorly from the perspective of a whole set may perform well in some of its parts, where alternative solutions fail, was the basis of the concept of creating ensemble of classifiers. Their theoretical advantage over the solutions based on individual classifiers was the assumed complementarity of models created with the use of various algorithms. The cost is to reduce the number of degrees of freedom.

There are many algorithms for creating ensembles that differ in their methods for obtaining the individual components and selection methods, and in the way in which the group identifies the probable class of an object. Often the result of voting of the majority of classifiers [4, 6–8], or another method resulting from the adopted combination of indications, is considered reliable. One possibility is the use of a "stacked generalisation" algorithm [9], performing the task of classification over several stages referred to as levels. The term *stacking* is found in the literature on the subject to refer to this type of classification construct. In the first stage of the classification process (called as "*level 0*"), a certain number of independently optimised classifiers (differing in their architecture, operating principles, etc.) interprets the original input data vector (covering the descriptive features of an object). Each of these classifiers generates its own class indication—the number of indications corresponds to the number of classifiers. In the final stage (*level 1*), a single classifier selects a class, taking into consideration the structure of the vector of indications in the previous processing stages. The classifier is optimised using a training set, comprising the first set of raw (original) data transformed by the components of the ensemble. The improvement in the degree of classification compared to individual classifiers depends on the structure of the group and the complexity of the task when individual classifiers of the first stage fail to identify some classes. This results from the formation of the specific structure of the vector of indications from the zero-level classifiers, forming a configuration linked by the first level with a specific, correctly determined class. A drawback of this approach based on a group of classifiers is the considerable increase in the time needed to achieve its final architecture. In addition to the optimisation of a certain number of classifiers (including ones that are redundant later in the process), transformation of the data set and optimisation of the interpreting classifier is required. The relationship between the benefits, related to the improvement of classification reliability, and the cost, including extra time and complications in data transformation, should be therefore studied. The aim of the studies presented is to determine the improvement in reliability of classification of soil complexes by introducing, in place of a single specialised classifier, a group built according to the *stacking* principle. The characteristics of the input model are particle size distribution (4 layers: 0–25, 25–75, 75–125, and >125 cm; content of clay and dust), the elements of morphology (ordinate, the slope of the surface, the configuration index k), and hydrological factors (depth ground water table, distance from watercourses and water bodies).

2. Description of the Problem of Classification

Agricultural soils in Poland are subject to several classifications with different purposes. The typological classification referring to the World Reference Base (systematics) is commonly known. It classifies soil as a natural object formed as a result of a specific genesis under specified conditions.

One important component of soil description is its valuation classification, the grouping of soils in units (valuation classes) of similar profitability. It is mainly used to establish a basis for taxation.

The division into complexes of agricultural suitability is a type of habitat classification. It is analysed in spatial planning and environmental protection activities, although its creation was motivated by the intention to support agricultural production with indications regarding selection of crops for specific soil conditions.

Individual classifications are interconnected, as are morphological and hydrological soil conditions [10–12]. Respective to changes in the morphology (e.g., due to mining works) and hydrology of a piece of land (e.g., due to regulation of water, construction of an underground water intake or occurrence of an obstacle in the terrain which limits flow), classification changes may also occur, including its typology, quality, and habitat classifications. In those cases, such transformations should be subject to prediction, regardless of their effects being revealed very slowly. The traditional methods for soil classification do not allow sufficiently precise prediction of changes, since they operate mainly within the soil's morphological criteria, which are secondary in relation to the topographical and hydrological factors. A reliable prediction of the effects of transformations requires a sufficiently large database to allow the provision of an essential set of rules linking the classification with the predicted components of the morphological, lithological, and hydrological situation of soils. It is necessary, however, to ensure representative source material to create an appropriate database, through which it is possible to find empirical rules linking the physiology of the terrain with classification units.

One of the ways to obtain the necessary information base is to utilise the available data on soil, morphology, and hydrology in order to empirically select a set of classification rules. The source of the bases of the rules may be the relationships between soil units and the terrain physiography included in cartographic and topographical soil documentation. To obtain such information, it is necessary to apply a data mining algorithm to derive rules from the database, whereby the database takes the form of a "black box."

For agricultural land in Poland, there are two types of soil and cartographic documentation. First, oil classification maps, a source document showing the diversity of agricultural land in Poland, were developed between 1956 and 1970 based on the field studies conducted nationwide. The main component of the 1 : 5000 scale classification maps are typological and quality classifications.

Second, the soil and agricultural maps originate from a soil classification map with content suitable for the needs of agriculture and spatial management. In addition to the classification (typological) content, it also includes spatial diversity of agricultural suitability complexes and composition of grains in the soil profile. An agricultural suitability complex is a soil and habitat unit indicating the specific suitability of a given parcel of land for specific crops based on its soil quality, moisture, location, and terrain morphology. The complex indicates overall morphological and soil conditions, as well as trends characterising the soil formation processes. Therefore, it indicates the proper management course through a variety of natural factors. It may be regarded as an indicator of

environmental conditions. In Poland, there are 14 agricultural suitability complexes for arable land (including three occurring in mountainous areas only), and three suitability complexes for permanent grassland [13]. The agricultural suitability complexes of soils comprise a nominal scale, in contrast to the quality classification. Identification of a complex through specific lithological, morphological, and hydrological conditions is a classification task.

2.1. Source of Data and Methodology. The source of data used for the research was a soil and agricultural topographic map of a part of the Upper Silesia Industrial Region with an area of 2596 square kilometres (rectangle dimensioned 59 km in the WE direction and 44 km in the NS direction). The Upper Silesia Industrial Region is varied in terms of morphology, hydrology, and soils. It is characterised by a developed hydrographical lattice. The soils have been transformed locally due to many years of mining activity. The soil and agricultural map indicates over 16,000 outlines of soil units in this area.

The typological diversity includes units of podsols, brown soils, black earth, hydrogenic soils, limestone soils, and alluvial soils. As far as quality is concerned, most of the quality classification units of the arable land occur (categories from II to IV), as well as a full range of lowland grassland. In addition to the most favourable wheat complex, all lowland arable and grassland complexes are listed there.

The database of soil diversification in the Upper Silesia Industrial Region was developed on the basis of a 1 : 25000 scale digitalised soil and agricultural map, and a topographic map on the same scale. The details of the soil contours were transformed in a regular lattice of adjacent squares of land with the area 400 m^2. Of 6,490,000 squares included in the study, a database of 2,767,890 squares was selected with complete information on soil and land morphology (soil complex, grain size distribution in the soil profile), the remaining area includes developed land or places in which there is no information on soils.

The database, in addition to the information derived directly from the soil, agricultural, and topographic map (square centre ordinate, complex identifier, soil type, grain size distribution, and inclination of the area), included contextual information: the k index, drainage area identifier, microdrainage area index, and the *Aer* and *Dist* indices.

The k index characterises the configuration of land around a square. Land configuration may be more favourable or less favourable to surface inflow. For a square in the Ith line of the jth column, the index is

$$k_{ij} = \beta \sum_{r=i-5}^{r=i+5} \sum_{k=j-5}^{k=j+5} \frac{z_{ij} - z_{rk}}{\left(x_{ij} - x_{rk}\right)^2 + \left(y_{ij} - y_{rk}\right)^2}, \quad (1)$$

where $(z_{ij} - z_{rk}) > 0$ and $i \neq r$ and $j \neq k$, $\beta = 1/120$.

This index determines whether the land configuration is favourable (higher value) or unfavourable (lower value) for water inflow from the adjacent area.

The drainage area identifier is a nominal variable indicating the parcel's inclusion in one of 46 drainage areas in this region (*r.waterhed* procedure in the GRASS package).

Table 1: Number of examples of individual complexes of agricultural suitability in data sets T-1, T-2, and V-1.

Complexes of agricultural suitability	Designation	T-1	T-2	V-1
Good wheat	K-2	5671	5848	5909
Defective wheat	K-3	2557	2619	2600
Very good rye	K-4	2591	2627	2571
Good rye	K-5	5040	4941	5068
Weak rye	K-6	5979	6235	6172
Very weak rye	K-7	1093	1051	1022
Strong grain and fodder	K-8	2061	2046	2021
Weak grain and fodder	K-9	1090	1101	1131
Very good and good permanent grassland	Z-1	20	22	20
Average permanent grassland	Z-2	5291	5316	5237
Very weak and weak permanent grassland	Z-3	1906	1826	1982

The microdrainage area index indicates the placement of the microdrainage area in an incremental sequence, ordered according to the ordinate of the lowest point of the area. Similarly, the square index in the microdrainage area is the placement in an incremental sequence, ordered according to squares belonging to the area. The *Aer* index is the average distance between the ground water surface and land area. The last contextual index is the distance of the square centre from the nearest edge of surface water (river or basin). The input data altogether included up to 16 variables, including 8 grain size indices in four layers of the soil profile (0–25 cm, 25–75 cm, 75–125 cm, and 125–150 cm). An element not included as an input variable (except for the complex label being the proposed model response) was the type of soil, which is closely related to the complex—it is, however, subject to changes in development trends that are difficult to predict resulting from changes to the hydrological situation. Typological prediction is a separate problem that is closely related to the problem of changes in quality classification or inclusion in agricultural suitability complexes.

2.2. Division of Data. The set of data related to agricultural land included 2,767,890 records. For the purpose of the study, 100,664 records were selected randomly. This set was, in turn, randomly divided into three parts, designated as T-1, T-2, and V-1. The distribution of samples belonging to individual agricultural suitability complexes was relatively uniform, except for the Z-1 complex, where the number of cases was very low (Table 1).

The T-1 set, including 33,299 records, was used to optimise 26 classifiers, from which the set that made up *level 0*, composed of 5 classifiers, was created. In accordance with the

prevailing optimisation rules for neural models, it was randomly divided into three approximately equal parts: training set of individual classifiers, a test set, and a validation set. The *Intelligent Problem Solver* procedure was applied, available in the *Statistica Neural Networks* package.

The T-2 set, made up of 33,632 records, was used to select the 5 classifiers with the most favourable properties from the initial 26, based on the evaluation of the number of correctly identified squares and the Q statistics calculated according to the formula [14]:

$$Q_{a,b} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}, \quad (2)$$

where N^{11} is the number of cases correctly classified by both classifiers, N^{00} is the number of cases incorrectly classified by both classifiers, N^{10} and N^{01} denote, respectively, the number of cases correctly classified by the first classifier and incorrectly by the other, and the number of cases with the inverse properties.

The Q statistics take a value 1.0 when the classifiers being compared show incorrect readings from the same cases. They take a negative value when the classifiers show incorrect readings for different cases. According to these properties, the group of classifiers with the lowest value of Q statistics will be most appropriate. The algorithm for selecting components from the set of classifiers from a group of 26 candidate classifiers involved inclusion of the strongest classifiers, and then inclusion of the following ones, characterised by the lowest average value of Q indices, in relation to the classifiers already present in the set.

The classifiers selected were then applied to process set T-2, converting it into set T-2C, composed of the vectors of readings of the classifiers in the set. With the labels of the correct classification of cases, they constituted a training set to optimise the classifier for *level 1*.

Set V-1, processed to become set V-1T, was a validating set for the classifier forming *level 1*, and also for the entire group. The data required prior processing analogous to that which converted set T-2 into set T-2C.

2.3. Classifiers Forming Level 0. The *Intelligent Problem Solver* (IPS) procedure in the *Statistica Neural Networks* package consists of creation and testing of many neural classifiers in order to determine the approximate location of the optimum point in the network architecture, matching the specified task, and an optimum set of input variables. The IPS parameters were the range of classifier architectures tested; number of layers hidden in the architectures requiring such information; the scope of the network dimension testing. In the optimisation of MLP architecture classifiers, reverse error propagation algorithms and reductions of the conjugated gradient were applied. The RBF classifiers were optimised using the *K-means* algorithm, with the aim of establishing the centres of radial function, and *K-nearest* neighbours was used to determine the scattering of the radial functions and pseudoinversions for the optimisation of the output layer weights. Probabilistic classifiers PNN were optimised by iteration, through changing the scattering of functions in the surrounding class patterns.

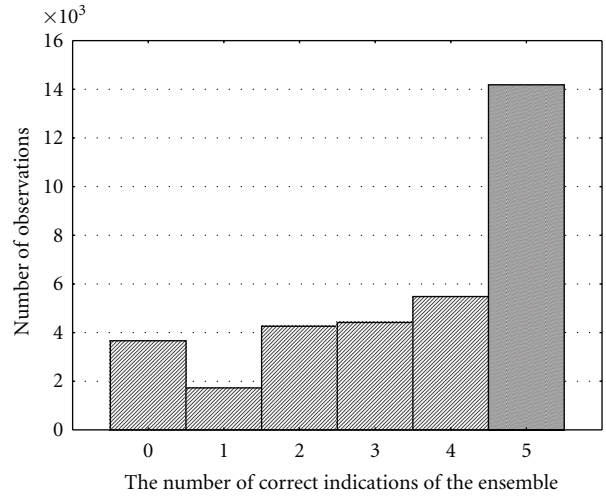

Figure 1: Histogram of the number of correct indications of the elements of set V-1 by the classifiers making up *level 0*.

In relative terms, the best properties for the possibility to identify the components of the training set were shown by probabilistic networks (Table 2). However, neither of the classifiers can be regarded as good enough for practical application in assessing soils. In the best case, a reading error of up to 30% of the analysed cases should be assumed. Presumably, improvement of the architecture, postprocessing parameters, and increasing the number of training cycles may lead to some slight improvement of the properties of individual classifiers, especially in physiographically homogeneous areas.

In the next step, based on the Q criterion, 5 classifiers performing the first stage of the classification (*level 0*) were selected. Table 2 shows some properties of the selected classifiers.

The quality of the set formed in this way may be analysed as a whole, assuming the voting rule as the classification criterion. Figure 1 shows the number of the V-1 validation set components, shown correctly, respectively, by sets of 0, 1, 2, 3, 4, and 5 classifiers. Use of three or more classifiers correctly indicated 24,080 (71.4%) of the elements of set V-1. The efficiency of a simple voting of the set in relation to individual classes was not as good in relative terms (Table 3).

Figure 2 shows the distribution of the correctness of individual complexes, the patterns of which were located in the set V-1. According to the figure, none of the agricultural suitability complexes was identified without errors. Another important finding is the very low identification effectiveness of the K-9, Z-1 and Z-2 complexes by this group. This can be regarded as a considerable drawback of the model, since these complexes are most vulnerable to hydrological changes.

2.4. Classifiers Forming Level 1. The classifier making up *level 1* can by any classifying neural network or other classifying algorithm. The obvious condition is its optimisation in relation to the inputs and labels of the set processed by the components of *level 0*. The inputs of the training set applied for optimisation are the vectors of indications of the *0-level*

Table 2: Some properties of classifiers obtained from the *Intelligent Problem Solver* procedure. Acronyms: MLP: Multilayer Perceptron, RBF: Radial Basis Function, PNN: Probabilistic Neural Network, TPerf: identification correctness of the training part of the training T-1 set, VPerf: identification correctness of the validating part of the T-1c set, TePerf: identification correctness of the test part of the T-1 set.

Type of classifiers	Number of inputs	Number of hidden units	Range of values TPerf	Range of values VPerf	Range of values TePerf
MLP	1–16	2–24	0.34–0.75	0.34–0.73	0.34–0.73
RBF	8–16	10–72	0.45–0.61	0.45–0.61	0.46–0.60
PNN	16	10000	0.53–0.93	0.51–0.72	0.53–0.72

Table 3: Some properties of the classifiers forming the level 0. Acronyms: MLP: Multilayer Perceptron, RBF: Radial Basis Function, PNN: Probabilistic Neural Network, other acronyms as in Table 2.

Typ	Inputs	Hidden	TPerf	VPerf	TePerf
MLP	6	6	0.6934	0.6872	0.696895
MLP	15	19	0.7429	0.725	0.732536
RBF	8	10	0.5517	0.5373	0.556959
PNN	16	10000	0.9925	0.7293	0.730431
PNN	16	10000	0.9375	0.7257	0.728701

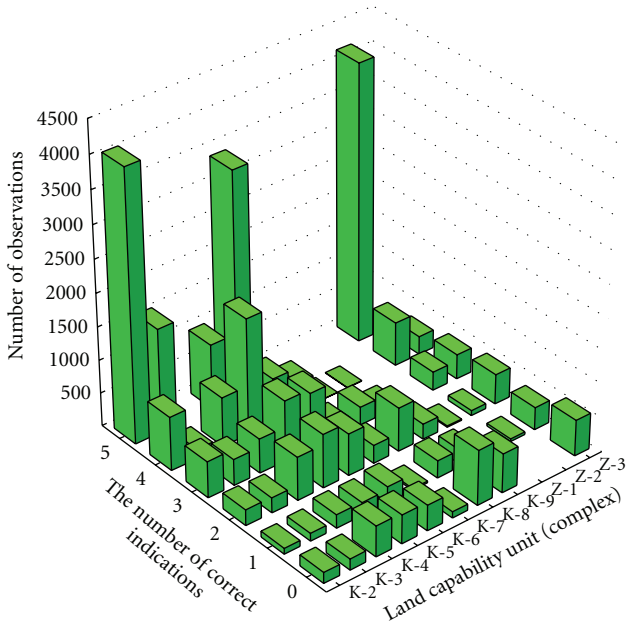

Figure 2: Distribution of the effectiveness of identification of complexes by a group of five classifiers forming the *level 0*.

classifiers, and the postulated responses are the correct labels of the individual teaching cases.

The best classifier for determining the ultimate indication of a set was sought in two groups:

(i) in a group of conventional neural networks using the algorithm of the input search (*Intelligent Problem Solver*) for an appropriate architecture within the multilayer perceptrons (MLP), networks with radial basic functions, and probabilistic networks (PNN);

(ii) in a group of constructivist algorithms, where the optimisation (number of hidden units, shape of the transfer function) is an element of training of a single network; the algorithms referred to as FSM and IncNet [4] were tested in these.

The construction of the MLP, RBF and PNN classifiers, which can be regarded as conventional, involves the creation of an appropriate prototype and its optimisation without changing the architecture. Systematic checks of classifiers with different sizes allow selection of the best model among those tested.

Classifiers formed in the constructivist procedures adjust their architecture to the complexity of the classification task in the optimisation process. This allows some reduction in the number of attempts required to construct an appropriate structure that is near the optimum ideal. Decisions as to the detailed structures of the classifiers (the transfer function form, initiation method and stop criterion) are, however, inevitable.

The IncNet, or *Incremental Network* [4, 6, 15], classifier is formed as a result of using an algorithm generating a group of networks with the RBP architecture, composed of single classifiers in the number corresponding to that of the identifiable classes (one classifier per class). The networks forming the IncNet groups differ in size and variables, optimising the differentiation of a specific, selected class of objects from other classes. Teaching the network during the implementation of the GhostMiner program involves the application of an extended Kalman filter. During training, the networks adjust their architecture by adding or removing RBF units to their respective functions in the identification process.

The FSM classifier, *Feature Space Mapping* [4], is an original ontogenic network solution with a structure corresponding to the neurofuzzy system. In the training process, there are stages of increase and decrease of the number of processing units to improve the consistency of the probability density distribution model with the observed distribution of features. The training of the network continues until the aim of optimisation is achieved, as expressed by the postulated effectiveness of identification.

3. Results

A relatively objective assessment of the correctness of the classification procedure is possible through the processing of the validation set (Tables 4 and 5).

Search for an optimum architecture in the scope of conventional neural networks included the MLP classifier (from 5 do 35 units in the hidden layer, logistical transfer function,

Table 4: Percent effectiveness of the identification of complexes of agricultural suitability of soils in validation set V-1 by selected classifiers making up *level 1* MLP, RBF, and PNN. VPerf: correctness of identification of complexes, VErr: complexes incorrectly included in the class.

Level 1	MLP		RBF		PNN	
Designation	VPerf	VErr	VPerf	VErr	VPerf	VErr
K-2	85.7	K-3 to K-8	84.9	K-3 to K-8	84.7	K-3 to K-8
K-3	85.3	K-2 to K-9	79.0	K-2 to K-9	76.1	K-2 to K-9
K-4	70.0	K-2 to K-9	67.4	K-2 to K-9	70.5	K-2 to K-9
K-5	81.9	K-2 to K-9	82.4	K-2 to K-9	84.3	K-2 to K-9
K-6	79.1	K-2 to K-9	81.7	K-2 to K-9	78.6	K-2 to K-9
K-7	82.6	K-2 to K-9	85.4	K-2 to K-9	87.0	K-2 to K-9
K-8	45.2	K-2 to K-6	45.3	K-2 to K-6	44.9	K-2 to K-6
K-9	27.1	K-2 to K-7	23.5	K-2 to K-7	33.6	K-2 to K-7
Z-1	0.0	Z-2	0.0	Z-2	100.0	
Z-2	89.6	Z-3	93.5	Z-3	90.6	Z-3
Z-3	65.0	Z-2	52.7	Z-2	62.0	Z-2

Table 5: Percent effectiveness of the identification of complexes of agricultural suitability of soils of validation set V-1 by selected classifiers forming *level 1*. IncNet, FSM composed of 25 units in the hidden layer, and FSM composed of 31 units in the hidden layer. VPerf: correctness of identification of complexes, VErr: complexes incorrectly included in the class.

Level 1	IncNet		FSM(25)		FSM(31)	
Designation	VPerf	VErr	VPerf	VErr	VPerf	VErr
K-2	100.0		100.0		100.0	
K-3	87.2	K-2, K-4	100.0		100.0	
K-4	100.0		100.0		100.0	
K-5	100.0		99.9	K-7	100.0	
K-6	100.0		99.9	K-7	100.0	
K-7	98.4	K-8	100.0		100.0	
K-8	93.3	K-7, K-9	100.0		100.0	
K-9	98.5	K-8	93.9	K-7, Z-2	100.0	
Z-1	0.0	Z-2	0.0	Z-2	100.0	
Z-2	100.0		100.0		99.9	Z-1
Z-3	100.0		94.2	Z-2	100.0	

minimisation of the average error as the optimisation criterion, BFGS—Broyden-Fletcher-Goldfarb-Shanno training algorithm, a version of the Quasi-Newton algorithm), RBF classifier (5–40 Gauss hidden units, pseudo-inverse optimisation for the output layer), and PNN probabilistic classifier (10,000 hidden units). The assessment of results obtained in this manner must be negative. Classification correctness of the best classifiers (Table 4) is similar (MLP—77.6%, RBF—77.1%, and PNN—77.3%). The error is too high, while the scope of redundancy indications (not belonging to the class) encompasses practically the entire scope of the classification. Within the classification of the individual classes, there are more or less significant errors; it is interesting that one of the classifiers (PNN) properly identifies one class which is not identified by any of the remaining ones.

Much better results are achieved when ontogenic algorithms are applied (Table 5).

Out of the 33,733 elements of set V-1, the IncNet algorithm properly classified 33,211 cases (522 errors), or 98.45% of the validation set. From an arithmetical point of view, this result is satisfactory; however, with the generally satisfying indications of individual complexes, the symptom of concern is the total absence of the possibility to isolate the Z-1 complex. The reason for this drawback is obvious: a very low proportion of this complex in the entire set (in the validation set, there are 20 cases, or just under 0.06%). Statistically, this share is unimportant, however, bearing in mind the morphological and hydrological features of this complex, it is difficult to accept the lack of possibility of a correct identification.

The results of using the FSM composed of 25 processing units in the hidden layer do not differ significantly from the results from the use of IncNet. The size of the hidden layer of this type of algorithm depends in this case on the complexity of the task and the stop criterion, which is the postulated identification effectiveness (99%). Similarly to the case of IncNet, the achieved effectiveness of identification of the validation set of elements is statistically very high (99.36%), which means that 33,518 elements are correctly identified

(only 215 cases are incorrectly classified). The decrease in the number of errors in relation to the IncNet network can be regarded as significant; however the Z-1 complex is still not correctly recognised.

Stricter stop conditions, at the level of 99.9% lead to a significant increase of units in the hidden layer, and a considerable improvement of the performance of the set as a whole. An ontogenic FSM algorithm with these stop conditions interrupted the construction of the network structure with an architecture composed of 31 units. Out of the 33,733 elements of validation set V-1, the Inc FSM(31) algorithm properly classified 33,727 cases (6 errors), which means 99.99% of the set. This result obtained for the validation set can be regarded as satisfactory. It is a characteristic that achievement of this relatively low, while important, improvement of the result required an increase in the number of units in the hidden layer by approximately 25%.

3.1. Model Application Example. In practice, there is a relatively limited number of cases in which considerable physiographic transformations may occur, which lead to changes in the classification of the soils. In prediction of such effects, the changes in soil development trends may be considered, because only prolonged action of the changed conditions (except for the catastrophic cases) will modify the habitats sufficiently for a transfer of one square of lad to other classification unit. The changes may accompany mining operations, leading to the lowering of ordinates of the land and its curvature, hydrographic, and hydrologic changes (draying or watering). Hydrological changes may accompany, to a lesser extent, the application of underground water intakes.

The area of the Upper Silesian Industrial Region has for many years been subjected to mining operations. This is accompanied by subsiding of the land, leading also to changes to its configuration. Figure 3 shows a part of a mining area of a mines in USIR. Part (a) of the figure shows the distribution of agricultural suitability complexes in this fragment, immediately before the beginning of mining operations. Part (b) of the figure shows designated squares of land in which changes to the development trends of agricultural habitats will occur, together with the direction of those changes. The changes in these trends impact a relatively small part of the land, where they appear mainly at the border of the soil complexes.

It is particularly visible in part (c) of the figure, where the predicted changes are visible mainly as the shifts in the contours of previously isolated classification units. This is an intuitive image of transitions, resulting, however, from the implicitly defined classification rules in the form of an empirical model. Due to relatively high reliability of the model, the predicted image of the future conditions of soils and habitats may be the basis for decisions regarding the significance of the influence of the operations on the environment, and also in planning and essential monitoring for documentation of the transitions.

4. Summary

Relationships between terrain physiography, usability, and classification of soils have long been known. The Jenny theory [10] is known, defining these relationships in a qualitative manner. The SCORPAN model [11] is based on similar assumptions, while it would be difficult to find its analytic form. Similarly, the model known as SoLIM [12] directly reflects the assumptions related to the relationships between physiography and classification of soils. In each of the cases mentioned, a certain prototype is available without representation in an analytic form. For that reason, accepting the fact of the relationship between physiography and soils, the model can be formed in an empirical procedure only, for example, using statistical methods. The statistics, however, require definition of the form of the relationships, that is, a regression equation prototype or a classification (usually linear) model. Its benefit is the possibility of analysing the variables obtained and the explicitness of their interrelationships. An alternative to this is computational intelligence methods, one version of which was applied in the research presented.

The claim of the lack of knowledge of the internal structure of the models, referred to as the "black box," which is difficult to challenge, has a significant importance here. Only a general knowledge of the processing rule, with the diversity of its parameters, prevents (hinders) the analysis of the problem solving method used by the model. In the recent years, models have been introduced which alleviate this shortcoming to some degree. Among others, the FSM algorithm, in rather limited conditions, makes it possible to define an appropriate set of rules in an *if…and/if…then…else* form for the classification. It is an important achievement, provided that the set of rules is sufficiently small. A set of several hundred rules may be processed effectively only using digital techniques. Therefore, the black box solution is an attractive option for a classification model, if validated properly. In the case analysed, a quite satisfactory classification result was obtained with a model composed of 5 classifiers for the first stage of processing and a defining classifier with 31 processing units. It must be emphasised that the results tested on an validation set independent from the training and test sets indicate stability of the model. It is also worth mentioning that it is suitable for an area of more than 2,500 square kilometres with a high level of physiographical diversification.

The considerable improvement of the credibility of the model with the use of a set of classifiers should also be emphasised. Considerable improvement of the effectiveness of the model resulted from the use of constructivist algorithms as a method for selecting the architectures for the second-stage classifier. It cannot be foreseen that in any possible case this procedure will be successful; however the main conclusion resulting from the research is that, in the classification procedures of the spatial information systems, an introduction of various model architectures would be desirable, with the possibility of combining them into sets with a flexible selection of components. It appears that the main reason for the advantage of the constructivist algorithms over the conventional MLP and RBP network training procedures is

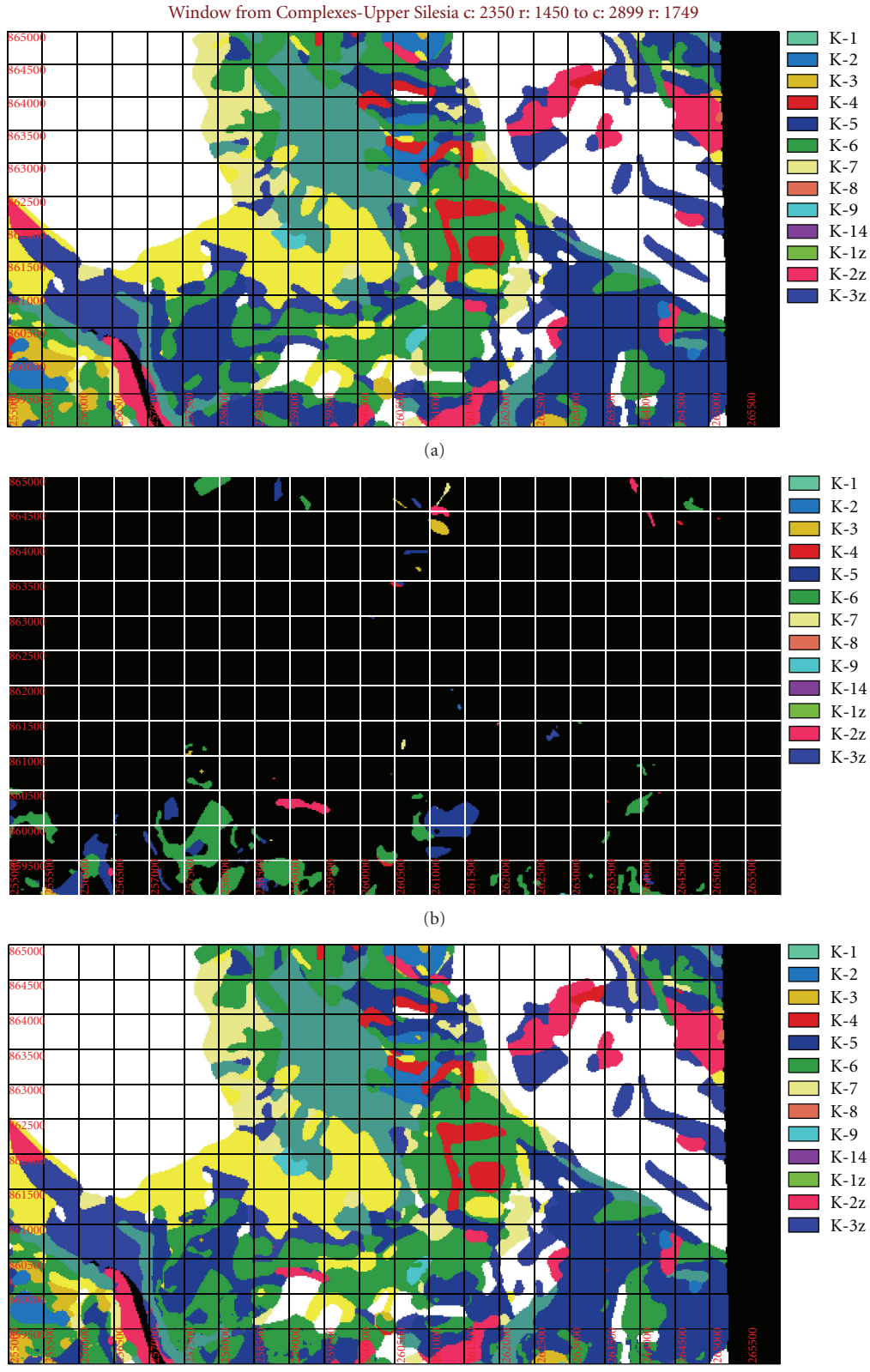

Figure 3: Prediction of complex change under the influence of mining operations in a part of the USIR area. (a) Spatial distribution of soil complexes before beginning of extraction, (b) land squares with changes causing considerable change of development trends of habitats, and (c) predicted distribution of the complexes of agricultural suitability.

an individual optimisation of the processing units, which adjusts the processing structure to the local complexity of the task, in contrast to the nonlocal optimisation in the conventional training procedures.

References

[1] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.

[2] R. Hecht-Nielsen, *Neurocomputing*, Addison-Wesley, Reading, UK, 1991.

[3] R. Tadeusiewicz, *Neural Networks*, Akademicka Oficyna Wydawnicza, Warszawa, Poland, 1993.

[4] N. Jankowski, *Ontogenic Neural Networks. The Networks Change their Structure*, Exit, Warszawa, Poland, 2004.

[5] J. Zurada, M. Barski, and W. Jędruch, *Artificial Neural Networks*, Wydawnictwo Naukowe PWN, Warszawa, Poland, 1996.

[6] W. Duch and K. Grabczewski, "Heterogeneous adaptive systems," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '02)*, pp. 524–529, May 2002.

[7] L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Wiley-Interscience, Hoboken, NJ, USA, 2004.

[8] S. K. Pal and P. Mitra, *Pattern Recognition Algorithms for Data Mining. Scalability, Knowledge Discovery and Soft Granular Computing*, Chapman and Hall/CRC Press Company, London, UK, 2004.

[9] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[10] H. Jenny, *Factors of Soil Formation. A System of Quantitative Pedology*, Dover Press, New York, NY, USA, 1994.

[11] A. B. McBratney, M. L. Mendonça Santos, and B. Minasny, "On digital soil mapping," *Geoderma*, vol. 117, no. 1-2, pp. 3–52, 2003.

[12] A. X. Zhu, "Mapping soil landscape as spatial continua: the neural network approach," *Water Resources Research*, vol. 36, no. 3, pp. 663–677, 2000.

[13] T. Witek, "The potential production capacity of arable land in Poland," *Roczniki Gleboznawcze*, vol. 36, no. 3, pp. 37–42, 1985 (Polish).

[14] M. Aksela, "Comparison of classifier selection methods for improving committee performance," in *Proceedings of the 4th International Conference on Multiple Classifier Systems (MCS '03)*, pp. 84–93, 2003.

[15] W. Duch, R. Setiono, and J. M. Zurada, "Computational intelligence methods for rule-based data understanding," *Proceedings of the IEEE*, vol. 92, no. 5, pp. 771–805, 2004.