UDC 001.891:65.011.56

THE GENERAL CONCEPT OF THE METHODS OF ALGORITHMIC CLASSIFICATION TREES

Povkhan I. F. – PhD, Associate Professor, Associate Professor at the Department of Software Systems, Uzhhorod National University, Uzhhorod, Ukraine.

ABSTRACT

Context. The general problem of constructing logical trees of recognition (classification) in the theory of artificial intelligence is considered in this paper. The object of this study is the concept of the classification tree (a logical and an algorithmic ones). The current methods and algorithms for constructing algorithmic classification trees are the subject of the study.

Objective. This work aims to create a simple and effective method for constructing tree-like recognition models on the basis of algorithmic classification trees for the training set of discrete information, which is characterized by the structure of the logical classification trees obtained on the basis of independent classification algorithms evaluated through the functional of calculating their overall efficiency.

Method. The general method of constructing algorithmic classification trees is proposed. It builds a tree-like structure (a classification model) for a given initial training data set. This structure consists of a set of autonomous algorithms of classification and recognition which have been evaluated at each step (stage) of constructing the model based on the given initial dataset. Namely, the method for constructing the algorithmic classification tree is proposed. The main idea of this method is to step by step approximate the initial dataset of arbitrary size and structure using a set of independent classification algorithms. This method, when forming the current vertex of the algorithmic tree (a node, a generalized feature) ensures the selection of the most effective (high-quality) autonomous classification algorithms from the initial dataset. In the process of constructing the resulting classification tree this approach can significantly reduce the size and complexity of the tree (the total number of branches, vertices and tiers of the structure) and improve the quality of its subsequent analysis (interpretability), the possibility of decomposition. The proposed method of constructing an algorithmic classification tree enables building different types of tree-like recognition models for a wide class of problems in the theory of artificial intelligence.

Results. The algorithmic classification tree method, developed and presented in this work, was implemented in the software and was studied and compared with the methods of logical classification trees (based on the selection of a set of elementary features) when solving the problem of recognizing real data of the geologic type.

Conclusions. The results of the conducted experiments described in this paper confirm the functional efficiency of the proposed mathematical software and show the possibility of its future use for solving a wide range of practical problems of recognition and classification. Further research prospects and approbation may consist in developing a limited method of the algorithmic classification tree, whose main points include the introduction of the criterion for stopping the procedure of constructing a tree model based on the depth of the structure, optimization of its software implementations, introduction of new types of algorithmic trees, and also the experimental research of this method while applying it for solving a wider range of practical problems.

KEYWORDS: algorithmic classification tree, pattern recognition, classification, classification algorithm, branching criterion.

ABBREVIATIONS

TS – training set;
CP – recognition system;
PR – pattern recognition;
GF – generalized feature;
RF – recognition function;
LCT – logical classification tree;
ACT – algorithmic classification tree;
GFT – generalized feature tree;
BFS – branch feature selection.

NOMENCLATURE

a_i is a fixed independent classification and recognition algorithms in the scheme of the algorithmic classification tree;

G is a some initial set of signals (discrete objects);

R is a partitioning into classes (patterns) H_i specified in the initial data set G;

f_R is a recognition function (RF) defined in the initial data set G;

x_i is a discrete objects (signals) of the initial TS;

H_i is a set of patterns (classes) specified in the initial TS;

$(x_i, f_R(x_i))$ is a training pairs of the initial TS;

m is a total number of training pairs (objects of the known classification) of the initial TS;

M is a total number of independent classification algorithms a_i in the set;

k is a total number of classes (patterns) of the set of signals G;

n is a total number of features (attributes) of a problem (feature space dimension);

l is a value of class membership of discrete object x;

$G_{a_1,...,a_i}$ is a subset of the initial G which is approximated using the set of classification algorithms a_i;

$S_{a_1,...,a_n}$ is a number of occurrences in the TS of training pairs which meet the condition $x_i \in G_{a_1,...,a_n}$;

$S_{a_1,\ldots,a_n}^{l}$ is a number of all training pairs of the initial TS which are correctly classified (when the basic membership requirement is fulfilled $x_s \in G_{a_1,\ldots,a_i}, f_R(x_s) = l$;

$\delta_{a_1,\ldots,a_i}$ is a frequency of occurrence of the members of the sequence of discrete objects from the initial TS in $G_{a_1,\ldots,a_i}$ data set;

$\psi_{a_1,\ldots,a_i}^{j}$ is a frequency of membership of object x in pattern (class) H_j provided that $x \in G_{a_1,\ldots,a_i}$;

$\rho_{a_1,\ldots,a_i}$ is a maximum frequency out of frequencies $\psi_{a_1,\ldots,a_i}^{j}$ regarding all the classes of the current problem;

F_S is a value which characterizes the quality of approximation of the initial TS using the set of classification algorithms a_i;

F is a function of recognizing the structure of the classification tree;

m_F is a number of all training pairs $(x_j, f_R(x_j))$ of the initial TS which are classified correctly;

τ_F is a value which characterizes the efficiency of RF for the initial TS regarding some set of classification algorithms $a_1,\ldots,a_i$;

f_j is a some generalized feature constructed at $j-$ step of generating the algorithmic classification tree model;

En_{tr} is a classification errors, rejections (of the first type) in the initial training data set;

Et_{tr} is a classification errors, rejections (of the first type) in the training data set;

T_{Uz} is a characterizes the total time (hardware time) which is necessary for building the current generalized feature f_j;

E_{Uz} is a information capacity (structural complexity) of the constructed GF f_j within the current step of generating the algorithmic classification tree model;

S_{Uz} is a total number of discrete objects x_i of the initial TS which are generalized (described) by the specified generalized feature f_j;

$P_{pt}(TS)$ is a power (volume) of the initial TS (or its fixed part for the current step of the scheme of constructing the algorithmic classification tree);

Er_{All} is a total number of errors of the algorithmic classification tree model in the data of the initial test set as well as of the training set;

M_{All} is a total power (volume) of the data of the training set as well as of the test set;

Fr_{All} is a characterizes the number of vertices of the obtained algorithmic classification tree model with the resulting value f_R (RF, i. e. the leaves of the classification tree);

V_{All} is a represents the total number of all types of vertices in the structure of the algorithmic classification tree model;

O_{Uz} is a total number of generalized features used in the classification tree model;

P_{All} is a total number of transitions between the vertices in the structure of the constructed classification tree model;

N_{All} is a total number of different classification algorithms that are used in the classification tree model;

I_{Main} is a indicator of generalizing data of the initial TS using the classification tree;

Q_{Main} is a integral indicator of the quality of the algorithmic classification tree model

INTRODUCTION

Today, the rapid scientific and technological advancements urge an engineer to solve the fundamentally important problem that often arises when working with large amounts of data. This is the problem of efficient automatic construction of systems for processing large amounts of information, decision-making systems, data analysis systems. It is clear that the solution of this problem allows one to completely transfer the hard work related to designing a complex system to a computer and release an engineer's creative potential to solve other, more important and relevant problems. Moreover, overcoming this problem within the theory of artificial intelligence along with automating the algorithm and software construction of specific recognition systems in the form of LCT / ACT models is the key to their high efficiency for every real problem, and consequently will ensure the rapid development of various fields of science and technology [1].

Information technologies based on mathematical models of pattern recognition in the form of tree models are widely used in systems of processing and analyzing arrays of information. Apparently, this is due to the fact that this approach eliminates a set of shortcomings typical of classical methods and achieves a fundamentally new result efficiently and rationally using the power of computer information systems [2].

Today we know almost four thousand recognition methods and algorithms which are based on various approaches and concepts, but they all have certain limitations in their use – accuracy, speed of operation, memory. It should be underlined that each of the existing classification algorithms is limited to certain specifics of application problems (universality constraints), and this is certainly the weakest point of not only these algorithms, but also information systems in general, which are based on the respective concepts [3, 4].

Decision trees, namely the structures of algorithmic classification trees are **the object of the study.**

Thus, the representation of training sets (arrays of discrete information) containing large amount of data in the form of structures of logical (algorithmic) trees has its significant advantages in terms of economic description and analysis of data, efficient mechanisms (procedures) for working with them [4, 5]. The coverage of the TS with the set of elementary features in the case with the LCT or the coverage of the TS with the fixed set of autonomous recognition and classification algorithms in the case with the ACT, generates a fixed tree-like data structure (a tree model) which provides even compression and conversion of the initial training dataset. This approach enables the significant optimization and saving of hardware resources of the information system [6–9].

It is known that the field of application of the concept of decision trees (LCT / ACT) is currently extremely extensive; yet many tasks and problems which are solved with the help of this instrument can be reduced to the following three basic segments – problems of describing data structures, recognition and classification problems, regression problems [10]. The vast majority of modern schemes of methods for constructing classification trees are known in the literature as 'divide and conquer'. It should be noted only that when this scheme is applied, the classification tree will be constructed in the direction from top to bottom [11].

An arbitrary structure of the classification tree (LCT / ACT) is presented in the form of branches and nodes, and on the branches of the tree there are some labels (attributes, values) on which the target function depends (in the case of the ACT – independent classification algorithms, sets of GFs), and the nodes (vertices) contain the RF values (the values of class membership) or the extended attributes of transitions.

The central issues of the concept of classification trees remain the ones linked to the choice of the branching criterion (constructing or selecting vertices), the branching stopping criterion (constructing the classification tree structure), and the criterion for rejecting branches of the logical tree (subtrees). This gives rise to the fundamental issue of the theory of classification trees which consists in the possible construction of all variants of logical trees that correspond to the initial TS and the selection of the minimal classification tree according to the depth, structural complexity (the number of tiers) [12–15].

The methods and algorithms for constructing algorithmic classification trees (decision trees) are **the subject of research**.

An important point about the existing methods of processing training sets (discrete arrays) in the recognition problems when classification rules are built is that they do not allow regulating their complexity (the parametric complexity of GFs) in the process of constructing the model [16–19]. This shortcoming is not found in the methods of constructing recognition systems which are based on the concept of algorithmic classification trees (decision trees). The peculiarity of the algorithmic tree method is the possibility of complex use for solving each specific problem of constructing the

scheme of recognition of many known recognition algorithms (methods). The concept of the ACT is based on the single methodology – the optimal approximation of the TS using the set of generalized features (autonomous algorithms), which are included in some scheme (operator), constructed in the training process [20].

The objective of the paper is to develop a simple, efficient and universal method of constructing models of classification (recognition) based on the concept of the ACT for discrete arrays, where the obtained schemes of classification systems are characterized by a tree-like structure and the presence of autonomous classification algorithms (the sets of GFs) as their structural elements.

1 PROBLEM STATEMENT

Suppose there is some initial TS as the sequence of training pairs which are the following:

$$(x_1, f_R(x_1)),...,(x_m, f_R(x_m)). \qquad (1)$$

Here the objects $x_i \in G$ (G is some set), and RF $f_R(x_i) \in \{0,1,2,...,k-1\}, (i=1,2,...,m)$. In addition to the initial TS, a test set (a set of objects of the known class membership) is also specified, as some part of the initial TS.

For RF $f_R(x_i) = l, (0 \le l \le k-1)$ means that $x_i \in H_l, H_l \subset G$, and here f_R is some finite significant function which specifies the initial partitioning R of G set which consists of subsets (a set of patterns, classes) $H_0, H_1, H_2,..., H_{k-1}$.

Hence, according to the specifications, a TS is a set (fixed sequence) of some sets (discrete objects), and each set is a set of values of some features (attributes) as well as values of some functions (RFs) typical of this set. Then the set of values of the features is a certain image (a discrete object), and the value of the function (RF) refers this image to the corresponding pattern [21–23].

Thus, the paper deals with the problem of constructing the ACT model with p parameters, whose L structure is optimal $F(L(p,x_i), f_R(x_i)) \to opt)$ in relation to the initial training dataset.

2 REVIEW OF THE LITERATURE

This research continues a series of works devoted to the problematics of tree-like recognition schemes (classification models) of discrete objects [4–7]. They highlight the issues of constructing, using and optimizing the structures of classification trees. As mentioned in [5], the resulting rule of classification (scheme), which is built with the help of an arbitrary method or algorithm of branched selection of features, has a tree-like logical structure, and the logical tree consists of vertices (features, attributes), which are grouped by tiers and which are obtained at a certain step (stage) of constructing the recognition tree [24]. An important task that arises from [20] is the one linked to synthesizing recognition

trees which will actually be represented by a tree (a graph) of algorithms (ACT methods). In contrast to the existing methods, the main feature of tree-like recognition systems is that the importance of individual features (groups of features or algorithms) is determined in reference to the function that specifies the partition of objects into classes [23]. Thus, work [15] is devoted to the principal issues concerning the generation of decision trees in the case of uninformative features, estimating the quality of the constructed models. The ability of classification tree structures to perform one-dimensional branching (the selection of features, attributes) for analyzing the impact (importance, quality) of individual variables (vertices) makes it possible to work with variables of different types in the form of predicates, generalized features, in the case of ACTs – with the respective autonomous classification and recognition algorithms. This concept of classification trees is actively used in data mining where the ultimate goal is to synthesize the model (fixed scheme), which predicts the value of the target variable based on the set of the initial data (training datasets) at the input of the system [26].

Nowadays a number of algorithms implementing the concept of decision trees (classification trees) are applied. However, their two representatives (C4.5 and CART) are widely spread and used; and also mentioned above the algorithm of the logical tree C4.5 as the node (vertex) selection criterion employs the so-called theoretical information criterion whereas CART algorithm is based on calculating Gini index which takes into account the relative distances (within the metric) between class distributions [27–30].

Since the main idea of methods and algorithms of the branched selection of features (vertices of algorithms) of ACTs can be defined as the optimal approximation of some initial TS using a set of ranked classification algorithms (features, attributes of the object in the case of LCTs), the central problem which arises is related to choosing an efficient branching criterion (the selection of vertices, attributes, features of discrete objects for LCTs and algorithms for ACTs). These fundamental problems are studied in [15, 21, 31–34]; they include quality evaluation of individual discrete features, their sets and fixed junctions; all of these enable introducing an efficient mechanism for the implementation of branching.

It should be noted that the structure of models of classification trees is characterized by compactness on the one hand and uneven occupancy (discharge) of the tiers on the other hand in comparison with regular trees [4]. It has to be stressed that the important issues, which still remain open, encompass the issue of the convergence of the process of constructing classification trees in accordance with the methods of branched feature selection and the issue of selecting the criterion for stopping the process of synthesizing a logical tree [19].

An important point is that there is no contradiction between the concept of classification trees and the possibility to use as the features (vertices of the structure) of the classification tree not only the individual attributes

(features) of objects of their junctions (the idea of the generalized feature which was studied in work [25]) and the sets; but if we go further and do not consider the attributes of objects (features) as branches – but select individual independent recognition algorithms (estimated based on the training dataset), a new structure – the ACT will be obtained at the output [20]. It is the ACT structures that this paper focuses on.

3 MATERIALS AND METHODS

By analogy with the methods of approximation of the TS using the set of estimated elementary features [3, 9, 15] – we present the main idea of the methods of algorithmic classification trees, which in turn lies in approximating the initial training dataset arrays with the help of the set of autonomous different-type algorithms of classification.

So, suppose there is the initial TS of general type (1) as a sequence of training pairs of the known classification (m power), and some system (set) of independent and autonomous recognition (classification) algorithms for the initial TS $a_1(x), a_2(x), ..., a_M(x)$. Then it is necessary to introduce the following sets which represent the partition of the training dataset using the respective classification algorithms a_i:

$$G_{a_1,...,a_i} = \{x \in G / a_i(x) = 1\}, (i = 1, ..., M). \quad (2)$$

It should be admitted that here in order to simplify the explanation each of the autonomous classification algorithms $a_i(x)$ generates at the output the values only within a binary set $\{0,1\}$, id est here $a_i(x) = 1$ in case of the successful classification of object x and $a_i(x) = 0$ in the opposite case.

Let us note that the system of $G_{a_1,...,a_i}$ sets actually represents the complete stepwise partitioning of G set (with the increasing i value – that is the involved classification algorithms), which is instantiated with the independent algorithms $a_1, a_2, ..., a_M$. It must be stressed that depending on the initial selection of the set of classification algorithms $a_1, a_2, ..., a_M$ some of the sets $G_{a_1,...,a_i}$ can be empty (the case of unsuitability of one specific or several algorithms for approximating the current TS).

At the next stage, by $S_{a_1,...,a_M}$ value we mark the number of occurrences in the initial TS of those training pairs $(x_s, f_R(x_s))$, $(1 \le s \le m)$ which meet the basic requirement of membership $x_s \in G_{a_1,...,a_i}$.

Respectively, by the value of $S_{a_1,...,a_M}^j$, $(j = 0, 1, ..., k-1)$ we mark the number of occurrences in the TS of those pairs $(x_s, f_R(x_s))$, $(s = 1, 2, ..., m)$ which meet the requirements $x_i \in G_{a_1,...,a_n}$ and $f_R(x_s) = j$.

Therefore, taking into account the above-mentioned and by analogy with the methods of selecting elementary feature sets, we can introduce the following values, which should be considered as a certain criterion of branching in the ACT structure:

$$\delta_{a_1,...,a_i} = \frac{s_{a_1,...,a_i}}{m}; \qquad (3a)$$

$$\psi_{a_1,...,a_i}^j = \frac{s_{a_1,...,a_i}^j}{s_{a_1,...,a_i}}; \qquad (3b)$$

$$\rho_{a_1,...,a_i} = \max_j \psi_{a_1,...,a_i}^j. \qquad (3c)$$

It should also be noted that if object $x_s \notin G_{a_1,...,a_i}$ for all $(s = 1,...,m)$, then it is clear that $\delta_{a_1,...,a_i} = 0$ and $\psi_{a_1,...,a_i}^j = 0$ when $(j = 0,1,...,k-1)$.

Hence, here $\delta_{a_1,...,a_i}$ value characterizes the frequency of occurrences of the members of the sequence $x_1, x_2,..., x_m$ (discrete objects) in $G_{a_1,...,a_i}$ set, and correspondingly $\psi_{a_1,...,a_i}^j$ value characterized the frequency of membership of the object of some x pattern (class) H_j if $x \in G_{a_1,...,a_i}$. It must be underscored here that the requirement $x \in G_{a_1,...,a_i}$ is equivalent to the requirement that in the sequence of algorithms $a_1,...,a_i$ there will be identified such an algorithm a_y in which $a_y(x) = 1$. Then the value $\delta_{a_1,...,a_i}$ characterizes information efficiency of recognizing the membership of some object x in one of classes $H_0, H_1,..., H_{k-1}$ certainly, provided that $x \in G_{a_1,...,a_i}$.

At the following stage again there arises the principal issue connected with the membership of object x in classes $H_0, H_1,..., H_{k-1}$ (the issue of the formation of classification rules). It is apparent that object x should also be referred to the class H_j for which the simple ratio is fulfilled:

$$\rho_{a_1,...,a_i} = \psi_{a_1,...,a_i}^j. \qquad (4)$$

We must stress that here $\{0 \le j \le k-1\}$, and the ratio (4) represents some classification rule, besides it is clear – the larger $\rho_{a_1,...,a_i}$ value, the higher its efficiency.

Since the initial TS is the only information which represents the partitioning of patterns $H_0, H_1,..., H_{k-1}$, and under H_j class we understand a set of all training pairs $(x_s, f_R(x_s))$ of the TS which meet the ratio $f_R(x_i) = j$, id est the requirement of membership. Similarly, the average efficiency of recognizing the set of patterns $H_0, H_1,..., H_{k-1}$ which are specified by the training

dataset using recognition algorithms $a_1, a_2,..., a_M$, is estimated by the following value:

$$F_S(a_1,...,a_M) = \sum_{a_1,...,a_i} \delta_{a_1,...,a_i} \cdot \rho_{a_1,...,a_i}. \qquad (5)$$

Hence, in this case $F_S(a_1,...,a_M)$ value can be calculated by means of the estimation of approximation of the initial TS using the set of independent classification algorithms $a_1, a_2,..., a_M$.

In view of the idea of the algorithmic classification tree which has been introduced above $F_S(a_1,...,a_M)$ value can be obtained on the basis of the following speculations – let us say that recognition function F is some image which puts some element of set $\{0,1,...,k-1\}$ (i. e., the corresponding number of the class) in line with each set $a_1,...,a_i$.

Thus, RF F is the function of type $F(a_1,...,a_i)$ where $a_1,...,a_i$ take values from $\{0,1\}$ set. According to RF $F(a_1,...,a_i)$ object $x,(x \in G)$ definitely refers to that class (pattern) $H_j,(0 \le j \le k-1)$ for which the following ratio is fulfilled:

$$F(a_1,...,a_i) = l,(0 \le l \le k-1). \qquad (6)$$

Suppose there is the initial TS of type (1), then let us consider that RF $F(a_1,...,a_i)$ classifies set $(x_j, f_R(x_j)),(1 \le j \le m)$ of the data array (1) correctly if $F(a_1(x_j),...,a_i(x_j)) = f_R(x_j)$, in the opposite case RF $F(a_1,...,a_i)$ classifies $(x_j, f_R(x_j))$ training pair incorrectly.

Next, suppose m_F – is the number of all occurrences of training pairs $(x_j, f_R(x_j))$ in the initial TS which classify RF $F(a_1,...,a_i)$ correctly. Then we introduce the following value:

$$\tau_F = \frac{m_F}{m}. \qquad (7)$$

It should be said that this value τ_F can be regarded as the total efficiency of RF $F(a_1,...,a_i)$ for the initial TS in reference to some set of classification algorithms $a_1,...,a_i$.

At the next stage of the research let us express this value τ_F through the previously proposed values $\delta_{a_1,...,a_i}$ and $\psi_{a_1,...,a_i}^j$. To do this let us count the quantity of those training pairs $(x_s, f_R(x_s))$ which are correctly classified by RF $F(a_1,...,a_i)$ and for which the ratio of membership $x_S \in G_{a_1,...,a_i}$ is fulfilled.

Suppose $F(a_1,...,a_i) = l$, then the number of all training pairs $(x_s, f_R(x_s))$ which are correctly classified (pro-

vided the basic requirement of membership is met $x_s \in G_{a_1,...,a_i}, f_R(x_s) = l$ is equal to $S^l_{a_1,...,a_i}$.

Then it becomes evident that m_F value, introduced above, is calculated according to the following formula:

$$m_F = \sum_{0 \le a_1,...,a_i \le 1} S^{F(a_1,...,a_i)}_{a_1,...,a_i}. \qquad (8)$$

It should be noted here that in this formula (8) we can take into account only those training sets for which the ratio $S_{a_1,...,a_i} \ne 0$ is fulfilled. Thus, in view of the comment concerning (8) that has just been indicated we can introduce m_F value in the following way:

$$m_F = \sum_{0 \le a_1,...,a_i \le 1} S_{a_1,...,a_i} \cdot \psi^{F(a_1,...,a_i)}_{a_1,...,a_i}. \qquad (9)$$

Let us admit that here $\psi^{F(a_1,...,a_i)}_{a_1,...,a_i}$ value was predefined in the expressions (4).

At the next stage the following is obtained from the formulae (9) and (8):

$$\tau_F = \frac{m_F}{m} = \sum_{0 \le a_1,...,a_i \le 1} \delta_{a_1,...,a_i} \cdot \psi^{F(a_1,...,a_i)}_{a_1,...,a_i}. \qquad (10)$$

It should be emphasized here that since $\psi^{F(a_1,...,a_i)}_{a_1,...,a_i} \le \rho_{a_1,...,a_i}$, then we shall have the following situation:

$$\tau_F \le F_S(a_1, a_2,...,a_M). \qquad (11)$$

Further, the recognition function (5) is denoted by F_0, and this function is specified by the ratio $F_0(a_1, a_2,...,a_n) = l$ if:

$$\psi^l_{a_1,...,a_i} = \rho_{a_1,...,a_i} = \max_{0 \le j \le k-1} \psi^l_{a_1,...,a_i}. \qquad (12)$$

Then based on the formulae (10) and (12) we immediately get the following:

$$\tau_{F_0} = \frac{m_{F_0}}{m} = \sum_{0 \le a_1,...,a_M \le 1} \delta_{a_1,...,a_M} \cdot \psi^{F_0(a_1,...,a_i)}_{a_1,...,a_i} =$$
$$= \sum_{0 \le a_1,...,a_M \le 1} \delta_{a_1,...,a_M} \cdot \rho_{a_1,...,a_M} = F(a_1,...,a_M). \qquad (13)$$

Hence, for all RFs $F(a_1,...,a_M)$ it follows from the formulae (13) and (11) that $\tau_F \le \tau_{F_0}$.

It should be admitted that the result of functioning of each of the fixed (selected from the library of algorithms of some information system) autonomous classification and recognition algorithms a_i at the respective step of generating the ACT is one or several generalized features f_j (certain classification rules) which, in fact, describe (approximate) the determined part of the initial training set. Thus, the respective resulting generalized features for

the case of the known geometric recognition algorithms [2] are geometric objects that cover the TS in feature space of the n dimension problem.

It is clear that in real examples there may be the cases when the respective classification algorithm a_i cannot construct the generalized feature f_j due to the complex arrangement of H_k classes in the feature space of the problem or specific conceptual and implementation constraints of the classification algorithm itself. Then, by analogy with the LCT such a case is possible when the generalized features f_j, constructed with the help of a_i classification algorithm, approximate the initial TS incompletely or such a situation is supported by the algorithm scheme of generating the ACT (as an example, the initial restriction in the algorithm scheme of the classification tree – about generating no more than one generalized feature f_j at each stage of constructing the ACT model).

It must be underlined that the objects of the initial TS, which do not fall under the constructed scheme of approximating the set using the sequence of generalized features f_j (at the final stage of the procedure of synthesizing the ACT), are referred to as rejections (errors) of classification of the first type – En_{tr} and similarly for the test dataset the incorrectly classified discrete objects are also referred to as errors of the first type – Et_{tr}.

Therefore, given all the above, we can assume that the ACT structure (type I) will have the general structure of the type – (Fig. 1), where each tier of such a classification tree determines the stage of constructing the ACT by means of approximating some part of the TS making use of the current classification algorithm a_i and owing to this approach enables adjusting the final complexity (accuracy) of the obtained classification tree model.

It should be stressed that within each step of generating the ACT model – (Fig. 1) there are given the specific classification algorithm a_i and respective TS (or the subset of the initial TS), and the initial TS in full is provided only at the first step, further with the subsequent steps of constructing a classification tree the power of the training dataset arrays will decrease due to the set of the constructed GFs f_j which will cut off (describe) some part of the initial training dataset. It is also important to emphasize that depending on the structure of the ACT construction scheme and the peculiarities of the current algorithm a_i more than one GF f_j can be generated within each step.

At the next stage of the research for the ACT method we introduce two basic criteria for constructing the classification tree model – the criterion for stopping the procedure of branching K_{stop} (it actually adjusts the complexity and accuracy of the obtained ACT model) and

the criterion for selecting branches $W(a)$ (the choice of a classification algorithm at the current step) for the classification tree under construction.

Thus, based on the above, it is advisable to introduce K_{stop} – the criterion for stopping the branching process of the type (*boolean*) of the ACT construction procedure which consists in checking the power $P_{pt}(TS)$ of the training dataset of the following type:

$$K_{Stop} = \begin{cases} 0, if\ P_{pt}(TS) = 0; \\ 1, if\ P_{pt}(TS) > 0. \end{cases} \quad (14)$$

Let us admit that the procedure of constructing a classification tree continues unless $K_{Stop} = 1$, and the opposite situation when $K_{Stop} = 0$ signals about finishing the stage of synthesizing the ACT model and the necessity of transition to the step of the test check based on the test dataset and the estimation of the quality of the obtained classification tree model.
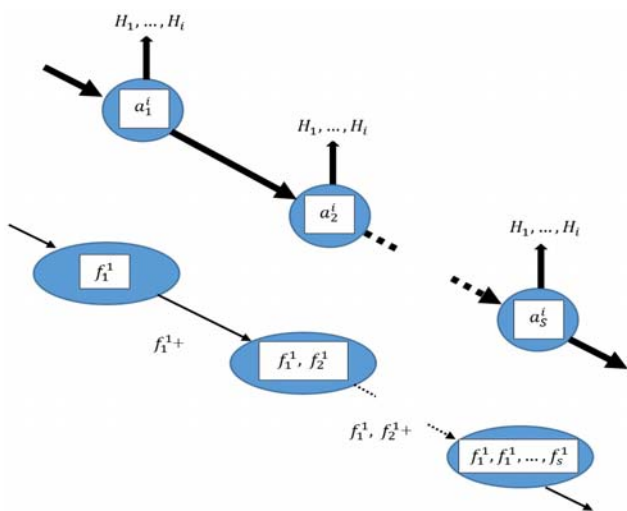

Figure 1 – The ACT structure of type I

It should be noted that while dealing with the ACT methods there arises the principal issue of selecting the branching criterion (the estimation and selection of classification algorithm a_i for the current step) in the structure of the classification tree model which is constructed. It is clear that similarly to the method of approximating the TS using the set of the ranked elementary features as the branching criterion we can suggest the initial estimation of efficiency of the set of algorithms $(a_1,...,a_i)$ of the following type:

$$W(a_i) = \frac{\frac{1}{P_{pt}(TS)} \cdot \sum_{j=1}^{k}(T_{Uz} + S_{Uz} + \frac{E_{Uz}}{S_{Uz}})}{k}. \quad (15)$$

Let us highlight that in the formula (15) summation is carried out based on all classes which are provided by the data array of the initial TS (though there can be restrictions on summation which are associated with the

structure (parameters) of the algorithm of constructing a classification tree).

An important point in the scheme of building the ACT model (Fig. 1) is that at each step of the tree algorithm actually its fixed (one or more – depending on the structure of the ACT algorithm itself) GF f_j is constructed, and along with this their total number increases with each coming step of the classification tree algorithm, namely the ACT with the set of classification algorithms $(a_1,...,a_m)$ generates (replicates) a tree-like structure – a tree of generalized features with the appropriate set of GFs $(f_1, f_2,..., f_z)$.

Therefore, in view of all the above, there can be offered one of the possible algorithmic schemes for constructing the ACT (type I).

Stage 1. At the first stage of constructing the ACT the set of autonomous classification and recognition algorithms $(a_1, a_2,...., a_M)$ is fixed in the library of algorithms of the information system (it is selected in an interactive mode randomly or after the procedure of appropriate estimation of the quality (efficiency) based on the initial training dataset). Let us note that both are selected – classification algorithms as well as their quantity (M value) depending on the aspects of the application problem.

Stage 2. At the second stage of constructing the ACT the selected set of classification algorithms $(a_1, a_2,...., a_M)$ is estimated and ranked on the basis of the functional (15) according to the training dataset in the set, respectively, with their efficiency. It should be stressed that here similarly to the LCT two options are possible – depending on the algorithmic scheme of constructing a classification tree:

a) The option when the estimation of efficiency and the ranking of the set of classification algorithms $(a_1, a_2,...., a_M)$ are done only once within this stage, and then at each step of constructing the ACT the following algorithm a_i of the initial sequence is fixed to approximate the data. This approach significantly saves the hardware resources of the information system, but negatively affects the complexity of the obtained classification tree model.

b) The option when the estimation of the efficiency and the ranking of the set of classification algorithms $(a_1, a_2,...., a_M)$ are done at each step of constructing the ACT according to the appropriate data of the subset (parts) of the initial TS in order to estimate and identify the highest quality (efficient) classification algorithm for this part of the TS (the step of generating the ACT). This approach enables completing the approximation of the TS in a fewer steps and obtaining a more economical structure of the ACT compared to option (a), however, it requires much more hardware resources of the information system for the second stage of the scheme of constructing the ACT and requires considerable attention and the introduction of a set of restrictions on the initial

selection of the set of classification algorithms $(a_1, a_2, ..., a_M)$.

Stage 3. At the third stage of constructing the ACT scheme the initial vertex of the ACT – the classification algorithm a_i of the highest efficiency out of the sorted set $(a_1, a_2, ..., a_M)$, is fixed, and at the input the initial TS in the form of the sequence of training pairs is provided. Algorithm a_i ensures the generating of one or several GFs f_i of the first tier (the number of GFs generated at each step are determined by the parameters of the algorithm of constructing the ACT) which approximate (classify) some part of the TS.

Stage 4. At the fourth stage the next most efficient classification algorithm a_i of the ranked sequence $(a_1, a_2, ..., a_M)$ is chosen as the vertex of the second tier and the procedure of constructing the GFs of the third stage is repeated with the only difference that at the input there is provided already the limited TS without training pairs which are approximated by the GF, the vertex of the first tier, etcetera. Thus, further the procedure of constructing the ACT will come down to repeating this stage for the next most efficient a_i algorithm sequence $(a_1, a_2, ..., a_M)$, the constant clipping of the parts of the TS and checking the branch stop criterion (the empty TS), which actually signals the completion of the procedure of constructing the ACT model and obtaining at the output a tree of classification algorithms a_i as well as a tree of generalized features f_i.

It should be mentioned that there are other implementation options of the scheme for constructing the ACT of the first type, which differ from the proposed scheme by variations in the number of GFs that are built at each step, the criteria and sequence of stages for assessing the quality of classification algorithms, the possibility of using a limited number of algorithms (even one), the possibility of approximating each of the classes of the TS using the set of its selected algorithms, the possibility of varying the criterion for stopping branching in the ACT.

Finally, it should be emphasized that the main peculiarity of such a scheme for constructing the ACT is the possibility to adjust the accuracy of the classification tree model, which is constructed during the basic procedure of constructing the ACT, along with this the important point is the principal possibility of constructing the ACT model with the predetermined accuracy with respect to the data array of the initial TS. This possibility is achieved by limiting the number of steps of the ACT generation procedure, the system of restrictions on information capacity, the number and parameters of generalization (the area of the approximated TS) of the set of generalized features which are constructed at the appropriate stages of constructing the resulting classification tree.

4 EXPERIMENTS

It has to be noted that the scheme for building the ACT suggested in this work enables adjusting the complexity of the classification tree model under construction, building models with the predetermined accuracy, and the classification tree structure consists of different-type autonomous classification algorithms as constructing modules (components). Moreover, the task of selecting the classification tree model from the set of constructed ACTs for a specific problem is determined by a set of parameters that are of decisive importance in relation to the current application problem (the training dataset). It is apparent that in order to compare and select a specific ACT model from the fixed set, it is necessary to determine their most important parameters (feature space dimension, the number of vertices, transitions, algorithms, etc.) and to identify their error regarding the input array.

It is fundamentally important to consider the criteria of quality of the obtained classification tree models, which depend on the model error, the power of the initial data array of the TS, the size of the test set (the number of training pairs and the dimension of the attribute space of the problem), the number of model parameters, etcetera [15, 27].

$$Q_{Main} = \frac{Fr_{All}}{V_{All} \cdot O_{Uz} \cdot P_{All} \cdot N_{All}} \cdot e^{-\frac{Er_{All}}{M_{All}}}. \quad (16)$$

Let us underline that this integral indicator of the quality of the ACT model will take values between zero and one. The lower it is, the worse the quality of the constructed classification tree is, and the higher the indicator, the better model is obtained.

An important indicator that characterizes the basic properties of the obtained ACT models is the general indicator of the generalization of the data of the initial TS using the classification tree; this indicator is calculated as follows:

$$I_{Main} = \frac{m \cdot O_{Uz}}{V_{All} + N_{All} + 2P_{All}}. \quad (17)$$

The proposed evaluation of the quality of the classification tree model (ACT) reflects its basic parameters (characteristics) of classification trees and can be applied as the criterion for optimality in the procedure of evaluating an arbitrary tree-like recognition scheme, for example in the case of methods of constructing and selecting random LCTs from work [24] (taking into account their respective structural parameters). It must be noted that the fundamental point which still remains is linked to reducing the complexity of the classification tree structure, the parameters of consuming memory λ and CPU time τ. It is also necessary to maximize the parameter I_{Main} (the indicator of generalization of the ACT model) that enables obtaining the most optimal structure of the classification tree and actually provides the maximum data compression of the initial TS (to represent the array of the initial data by the logical tree whose structural complexity is minimal) [12].

In Uzhhorod National University there has been developed "Orion III" software complex for generating autonomous recognition systems where the algorithmic library of the system has 11 recognition algorithms, among which there is the above-mentioned algorithmic implementation of constructing the ACT. Its task was to build an autonomous recognition system based on geological data (the problem of separating oil-bearing strata). 22 elementary features were used to identify the objects. The TS contains information about the objects of the two classes, and at the stage of the examination the constructed classification system (the ACT model) should ensure the effective recognition of objects of the unknown classification regarding these two classes. Note that before starting work, the training set was automatically checked for correctness (search and removal of the same objects of different membership – errors of the first kind)

5 RESULTS

Thus, the TS contains information on partition into two classes, and in the training data array there were predominantly training pairs of the first class (oil-bearing strata) in proportion (1.5 / 1), and the TS itself consisted of 1250 objects, and the efficiency of the constructed recognition system (the ACT model) was evaluated using the test set of 240 objects. The test set was a separate part of the training set (consisting of discrete objects of the known classification). The arrays of the training and test sets were obtained on the basis of geological exploration in Transcarpathian region during the period from 2001 to 2011.

Hence, a fragment of the main results of the above experiments (the comparative tests of methods for constructing LCT / ACT models using the data array of this application problem) is presented in (Table 1), and along with this the synthesized models of different-type classification trees provided the required level of accuracy specified in the task, the processing speed and costs of working memory of the information system, but showed different structural complexity of the constructed classification trees (models) and the set of generalized features in the case of the algorithmic classification tree model.

6 DISCUSSION

It is worth highlighting that the proposed estimates of the quality of the ACT model fix the most important characteristics of the classification trees and can be used as the criterion for optimality in the procedure of constructing the ACT and final selection from the set of ACT models. Notably, the method of the algorithmic tree operates only with the ready (constructed) generalized features, and it does not matter at all by means of what algorithm (rule, method) they were obtained, and each of the schemes constructed with the help of the algorithmic tree method is the general recognition system (the ATC model), which can be employed for practical work (processing large amounts of experimental data in the form of discrete sets). It should also be stressed that the resulting scheme is to some extent a new recognition algorithm (certainly, the one which has been synthesized from the known algorithms and methods). It is necessary to admit that the obtained ACT structure is characterized by the high degree of universality and relatively compact structure of the model itself, but it requires large hardware costs to store the generalized features of the initial assessment of the quality of classification algorithms according to the TS.

CONCLUSIONS

The problem of automation of constructing the models of algorithmic classification trees on the basis of approximating the TS using the set of independent classification algorithms has been solved in this paper.

The scientific novelty of the obtained results lies in the fact that for the first time there has been suggested a simple method of constructing the ACT based on evaluating and ranking the set of autonomous recognition algorithms for generating the classification tree structure (the ACT model). Moreover, within each step of branching the classification tree, a certain part of the TS (or its subset) is approximated. Admittedly, the functional (branching criterion) proposed in the work can be used not only for estimating the quality of individual classification algorithms, but also calculating the efficiency of the related sets of algorithms, which in the long run enables obtaining a more optimal structure of the synthesized ACT based on the initial training dataset.

Table 1 – The comparative table of the models / methods of classification trees.

No	The method of synthesizing the logical tree structure	Integral indicator of the model quality Q_{Main}	The total number of errors of the model in the training dataset and test dataset Er_{All}
1	The complete LCT method based on the selection of elementary features BFS-I	0.004691	2
2	The LCT method with one-time estimation of feature importance BFS-II	0.002172	3
3	The constrained BFS – LCT construction method	0.003219	4
4	The algorithmic tree method (type I)	0.005234	0
5	The algorithmic tree method (type II)	0.002941	0

The paper offers the set of general indicators (parameters) which allows the efficient presentation of the general characteristics of the ACT model, its possible use for selecting the most optimal ACT.

The applied value of the obtained results is that the proposed method of constructing ACT models was implemented in the library of algorithms of the universal software system "ORION III" to solve various practical problems of classification (recognition) of arrays of discrete objects. It must be underscored that the conducted practical tests confirm the functionality of the proposed ACT models and the developed software, which enables making a recommendation on the use of this approach and its software implementation for a wide range of application problems of classifying and recognizing discrete objects.

Further research is needed to develop methods of algorithmic classification trees (constrained methods) of optimizing software implementations of the proposed method for constructing the ACT as well as its practical approbation dealing with a number of real problems of classification and recognition.

REFERENCES

1. Srikant R., Agrawal R. Mining generalized association rules, *Future Generation Computer Systems,* 1997, Vol. 13, No. 2, pp. 161–180.
2. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Stanford, 2008, 768 p.
3. Quinlan J.R. Induction of Decision Trees, *Machine Learning*, 1986, No. 1, pp. 81–106.
4. Vasilenko Y. A., Vasilenko E. Y., Kuhayivsky A. I., Papp I. O. Construction and optimization of recongnizing systems, *Scientific and technical journal "Information technologies and systems",* 1999, No. 1, pp. 122–125.
5. Povhan I. Designing of recognition system of discrete objects, 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), 2016, Lviv, Ukraine. Lviv, 2016, pp. 226–231.
6. Mitchell T. Machine learning. New York, McGraw-Hill, 1997, 432 p.
7. Povhan I. General scheme for constructing the most complex logical tree of classification in pattern recognition discrete objects, *Collection of proceedings «Electronics and information technology»,* 2019, Vol. 11, pp. 73–80.
8. Breiman L. L., Friedman J. H., Olshen R. A., Stone C. J. Classification and regression trees. Boca Raton, Chapman and Hall/CRC, 1984, 368 p.
9. Vasilenko Y. A., Vashuk F. G., Povkhan I. F. Automating the construction of classification systems based on agent – schemes, *Mathematical modeling, optimization and information technologies : International Joint Conference MDIF-2012.* Kisheneu, Moldova, 2012, Kisheneu, 2012, pp. 444–446.
10. Vtogoff P. E. Incremental Induction of Decision Trees, *Machine Learning*, 1989, № 4, pp. 161−186.
11. Amit Y., Geman D., Wilder K. Joint induction of shape features and tree classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1997, Vol. 19, No. 11, pp. 1300–1305.
12. Dietterich T. G., Kong E. B. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms [Electronic resource]. Corvallis, Oregon State University, 1995, 14 p. Access mode : http://www.cems.uwe.ac.uk/~irjohnso/coursenotes/uqc832/trbias.pdf
13. Mingers J. An empirical comparison of pruning methods for decision tree induction, *Machine learning*, 1989, Vol. 4, No. 2, pp. 227–243.
14. Povhan I. Question of the optimality criterion of a regular logical tree based on the concept of similarity*, Collection of proceedings "Electronics and information technology",* 2020, Vol. 13, pp. 12–16.
15. Subbotin S.A. Construction of decision trees for the case of low-information features, *Radio Electronics, Computer Science, Control,* 2019, № 1, pp. 121–130.
16. Lupei M., Mitsa A., Repariuk V., Sharkan V. Identification of authorship of Ukrainian-language texts of journalistic style using neural networks, *Eastern-European Journal of Enterprise Technologies*, 2020, Vol. 1 (2 (103)), pp. 30–36. DOI: https://doi.org/10.15587/1729-4061.2020.195041
17. Bodyanskiy Y., Vynokurova O., Setlak G. and Pliss I. Hybrid neuro-neo-fuzzy system and its adaptive learning algorithm, *Computer Sciences and Information Technologies (CSIT) : Xth Scien. and Tech. Conf.* Lviv, 2015. Lviv, 2015, P. 111–114.
18. Karimi K., Hamilton H. J. Generation and Interpretation of Temporal Decision Rules, *International Journal of Computer Information Systems and Industrial Management Applications*, 2011, Vol. 3, pp. 314–323.
19. Kotsiantis S. B. Supervised Machine Learning: A Review of Classification Techniques, *Informatica,* 2007, No. 31, pp. 249–268.
20. Povhan I. F. Features of synthesis of generalized features in the construction of recognition systems using the logical tree method, *Information technologies and computer modeling ITKM-2019 : materials of the international scientific and practical conference, Ivano-Frankivsk, May 20–25, 2019.* Ivano-Frankivsk, 2019, pp. 169–174.
21. Vasilenko Y. A., Vashuk F. G., Povkhan I. F. The importance of discrete signs, *XX International Conference Promising ways and directions of improving the educational system, Uzhgorod, November 16–19, 2010.* Uzhgorod, 2010, Vol. 21, No. 1, pp. 217–222.
22. Deng H., Runger G., Tuv E. Bias of importance measures for multi-valued attributes and solutions, *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN).* Espoo, Finland, Jun 14–Jun 17, 2011, Espoo, 2011, pp. 293–300.
23. Kamiński B., Jakubczyk M., Szufel P. A framework for sensitivity analysis of decision trees, *Central European Journal of Operations Research,* 2017, Vol. 26 (1), pp. 135–159.
24. Dietterich T. G. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, *Machine learning*, 2000, Vol. 40, № 2, pp. 139–157.

25. Povhan I. Generation of elementary signs in the general scheme of the recognition system based on the logical tree, *Collection of proceedings «Electronics and information technology»*, 2019, Vol. 12, pp. 20–29.
26. Subbotin S., Oliinyk A. eds. : Szewczyk R., Kaliczyńska M. The dimensionality reduction methods based on computational intelligence in problems of object classification and diagnosis, *Recent Advances in Systems, Control and Information Technology*. Cham, Springer, 2017, pp. 11–19. (Advances in Intelligent Systems and Computing, vol. 543).
27. Subbotin S. A. Methods and characteristics of localitypreserving transformations in the problems of computational intelligence, *Radio Electronics, Computer Science, Control*, 2014, No. 1, pp. 120–128.
28. Koskimaki H., Juutilainen I., Laurinen P., Roning J. Two-level clustering approach to training data instance selection: a case study for the steel industry, *Neural Networks : International Joint Conference (IJCNN-2008)*. Hong Kong, 1–8 June 2008 : proceedings. Los Alamitos, IEEE, 2008, pp. 3044–3049. DOI: 10.1109/ijcnn.2008.4634228
29. Subbotin S. The neuro-fuzzy network synthesis and simplification on precedents in problems of diagnosis and pattern recognition, *Optical Memory and Neural Networks (Information Optics)*, 2013, Vol. 22, № 2, pp. 97–103. DOI: 10.3103/s1060992x13020082
30. Subbotin S.A. Methods of sampling based on exhaustive and evolutionary search, *Automatic Control and Computer Sciences*, 2013, Vol. 47, № 3, pp. 113–121. DOI: 10.3103/s0146411613030073
31. De Mántaras R. L. A distance-based attribute selection measure for decision tree induction, *Machine learning*, 1991, Vol. 6, № 1, pp. 81–92.
32. Alpaydin E. Introduction to Machine Learning. London, The MIT Press. 2010, 400 p.
33. Painsky A., Rosset S. Cross-validated variable selection in tree-based methods improves predictive performance, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, Vol. 39, No. 11, pp. 2142–2153. DOI:10.1109/tpami.2016.2636831
34. Miyakawa M. Criteria for selecting a variable in the construction of efficient decision trees, *IEEE Transactions on Computers*, 1989, Vol. 38, No. 1, pp. 130–141.

УДК 001.891:65.011.56

ЗАГАЛЬНА КОНЦЕПЦІЯ МЕТОДІВ АЛГОРИТМІЧНИХ ДЕРЕВ КЛАСИФІКАЦІЇ

Повхан І. Ф. – канд. техн. наук, доцент, доцент кафедри програмного забезпечення систем ДВНЗ Ужгородський національний університет, м. Ужгород, Україна.

АНОТАЦІЯ

Актуальність. Розглянута загальна задача побудови логічних дерев розпізнавання (класифікації) в теорії штучного інтелекту. Об'єктом даного дослідження є концепція дерева класифікації (логічного та алгоритмічного). Предметом дослідження є актуальні методи та алгоритми побудови алгоритмічних дерев класифікації.

Мета. Метою даної роботи є створення простого та ефективного методу побудови деревоподібних моделей розпізнавання на основі алгоритмічних дерев класифікації для навчальних вибірок дискретної інформації, який характеризується структурою отриманих логічних дерев класифікації з незалежних алгоритмів класифікації оцінених на основі функціоналу розрахунку їх загальної ефективності.

Метод. Пропонується загальний метод побудови алгоритмічних дерев класифікації, який для заданої початкової навчальної вибірки будує деревоподібну структуру (модель класифікації), яка складається з набору автономних алгоритмів класифікації та розпізнавання оцінених на кожному кроці (етапі) побудови моделі за даною початковою вибіркою. Тобто пропонується метод побудови алгоритмічного дерева класифікації основна ідея якого полягає в по кроковій апроксимації начальної вибірки довільного об'єму та структури набором незалежних алгоритмів класифікації. Даний метод при формуванні поточної вершини алгоритмічного дерева (вузла, узагальненої ознаки) забезпечує виділення найбільш ефективних (якісних) автономних алгоритмів класифікації з початкового набору. Такий підхід при побудові результуючого дерева класифікації дозволяє значно скоротити розмір та складність дерева (загальну кількість гілок, вершин та ярусів структури) підвищити якість його наступного аналізу (інтерпретабельність), можливість декомпозиції. Запропонований метод побудови алгоритмічного дерева класифікації дозволяє будувати різнотипні деревоподібні моделі розпізнавання для широкого класу задач теорії штучного інтелекту.

Результати. Розроблений та представлений в даній роботі метод алгоритмічного дерева класифікації отримав програмну реалізацію та був досліджений та порівняний з методами логічних дерев класифікації (на основі селекції набору елементарних ознак) при розв'язку задачі розпізнавання реальних даних геологічного типу.

Висновки. Проведені в даній роботі експерименти підтвердили працездатність запропонованого математичного забезпечення та показують можливість його перспективного використання для розв'язку широкого спектру практичних задач розпізнавання та класифікації. Перспективи подальших досліджень та апробацій можуть полягати в створенні обмеженого методу алгоритмічного дерева класифікації, який полягає в веденні критерію зупинки процедури побудови моделі дерева за глибиною структури, оптимізації його програмних реалізацій, введення нових типів алгоритмічних дерев а також експериментальних дослідженнях даного методу на більш широке коло практичних задач.

КЛЮЧОВІ СЛОВА: алгоритмічне дерево класифікації, розпізнавання образів, класифікація, алгоритм класифікації, критерій розгалуження.

УДК 001.891:65.011.56

ОБЩАЯ КОНЦЕПЦИЯ МЕТОДОВ АЛГОРИТМИЧЕСКИХ ДЕРЕВЬЕВ КЛАССИФИКАЦИИ

Повхан И. Ф. – канд. техн. наук, доцент, доцент кафедры программного обеспечения систем ГВУЗ Ужгородский национальный университет, г. Ужгород, Украина.

АННОТАЦИЯ

Актуальность. Рассмотрена общая задача построения логических деревьев распознавания (классификации) в теории искусственного интеллекта. Объектом данного исследования является концепция дерева классификации (логического и алгоритмического). Предметом исследования являются актуальные методы и алгоритмы построения алгоритмических деревьев классификации.

Цель. Целью данной работы является создание простого и эффективного метода построения древовидных моделей распознавания на основе алгоритмических деревьев классификации для учебных выборок дискретной информации, который характеризуется структурой полученных деревьев классификации из независимых алгоритмов классификации оцененных на основе функционала расчета их общей эффективности.

Метод. Предлагается общий метод построения алгоритмических деревьев классификации, который для заданной начальной учебной выборки строит древовидную структуру (модель классификации), которая состоит из набора автономных алгоритмов классификации, оцененных на каждом шагу (этапе) построения модели по данной начальной выборке. То есть предлагается метод построения алгоритмического дерева классификации основная идея которого заключается в по шаговой аппроксимации начальной выборки фиксированного объема и структуры набором независимых алгоритмов классификации. Данный метод при формировании текущей вершины дерева классификации (узла, обобщенного признака) обеспечивает выделение наиболее эффективных (качественных) автономных алгоритмов классификации из начального набора. Такой подход при построении результирующего дерева классификации позволяет значительно сократить размер и сложность дерева (общее количество ветвей, вершин и ярусов структуры) повысить качество его последующего анализа (интерпретабельность), возможность декомпозиции. Предложенный метод построения алгоритмического дерева классификации позволяет строить разнотипные древовидные модели распознавания для широкого класса задач теории искусственного интеллекта.

Результаты. Разработанный и представленный в данной работе метод алгоритмического дерева классификации получил программную реализацию и был исследован и сравнен с методами логических деревьев классификации (на основе селекции набора элементарных признаков) при решении задачи распознавания реальных данных геологического типа.

Выводы. Проведенные в данной работе эксперименты подтвердили работоспособность предложенного математического обеспечения и показывают возможность его перспективного использования для решения широкого спектра практических задач распознавания и классификации. Перспективы дальнейших исследований и апробаций могут заключаться в создании ограниченного алгоритмического метода дерева классификации, который заключается в ведении критерия остановки процедуры построения модели дерева по глубине структуры, оптимизации его программных реализаций, введение новых типов алгоритмических деревьев, а также экспериментальных исследованиях данного метода в более широкой области практических задач.

КЛЮЧЕВЫЕ СЛОВА: алгоритмическое дерево классификации, распознавание образов, классификация, алгоритм классификации, критерий разветвления.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Srikant R. Mining generalized association rules / R. Srikant, R. Agrawal // Future Generation Computer Systems. – 1997. – Vol. 13, № 2. – P. 161–180.
2. Hastie T. The Elements of Statistical Learning / T. Hastie, R. Tibshirani, J. Friedman. – Stanford, 2008. – 768 p.
3. Quinlan J. R. Induction of Decision Trees / J. R. Quinlan // Machine Learning. – 1986. − № 1. – P. 81–106.
4. Construction and optimization of recongnizing systems / [Y. A. Vasilenko, E. Y. Vasilenko, A. I. Kuhayivsky, I. O. Papp] // Scientific and technical journal "Information technologies and systems". – 1999. – № 1. – P. 122–125.
5. Povhan I. Designing of recognition system of discrete objects / I. Povhan // 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, 2016, Ukraine. – Lviv, 2016. – P. 226–231.
6. Mitchell T. Machine learning / T. Mitchell. – New York : McGrawHill, 1997. – 432 p.
7. Povhan I. General scheme for constructing the most complex logical tree of classification in pattern recognition discrete objects / I. Povhan // Collection of proceedings "Electronics and information technology". – 2019. – Vol. 11. – P. 73–80.
8. Classification and regression trees / L. L. Breiman, J. H. Friedman, R. A. Olshen et al.]. – Boca Raton : Chapman and Hall/CRC, 1984. – 368 p.
9. Vasilenko Y. A. Automating the construction of classification systems based on agent – schemes / Y. A. Vasilenko, F. G. Vashuk, I. F. Povhan // Mathematical modeling, optimization and information technologies : International Joint Conference MDIF-2012, Kisheneu, Moldova, 2012. – Kisheneu, 2012. – P. 444–446.
10. Vtogoff P. E. Incremental Induction of Decision Trees / P. E. Vtogoff // Machine Learning. – 1989. – №4. – P. 161−186.
11. Amit Y. Joint induction of shape features and tree classifiers / Y. Amit, D. Geman, K. Wilder // IEEE

Transactions on Pattern Analysis and Machine Intelligence. – 1997. – Vol. 19, № 11. – P. 1300–1305.

12. Dietterich T. G. Machine learning bias, statistical bias and statistical variance of decision tree algorithms [Electronic resource] / T. G. Dietterich, E. B. Kong. – Corvallis : Oregon State University, 1995. – 14 p. Access mode : http://www.cems.uwe.ac.uk/~irjohnso/coursenotes/uqc832/trbias.pdf

13. Mingers J. An empirical comparison of pruning methods for decision tree induction / J. Mingers // Machine learning. – 1989. – Vol. 4, № 2. – P. 227–243.

14. Povhan I. Question of the optimality criterion of a regular logical tree based on the concept of similarity / I. Povhan // Collection of proceedings «Electronics and information technology». – 2020. – Vol. 13. – P. 12–16.

15. Subbotin S. A. Construction of decision trees for the case of low-information features / S. A. Subbotin // Radio Electronics, Computer Science, Control. – 2019. – № 1. – P. 121–130.

16. Identification of authorship of Ukrainian-language texts of journalistic style using neural networks / [M. Lupei, A. Mitsa, V. Repariuk, V. Sharkan] // Eastern-European Journal of Enterprise Technologies. – 2020. – Vol. 1 (2 (103)). – P. 30–36. DOI: https://doi.org/10.15587/1729-4061.2020.195041

17. Hybrid neuro-neo-fuzzy system and its adaptive learning algorithm / [Y. Bodyanskiy, O. Vynokurova, G. Setlak and I. Pliss] // Computer Sciences and Information Technologies (CSIT) : Xth Scien. and Tech. Conf., Lviv, 2015. – Lviv, 2015. – P. 111–114.

18. Karimi K. Generatio n and Interpretation of Temporal Decision Rules / K. Karimi, H. J. Hamilton // International Journal of Computer Information Systems and Industrial Management Applications. – 2011. – Vol. 3. – P. 314–323.

19. Kotsiantis S. B. Supervised Machine Learning: A Review of Classification Techniques / S. B. Kotsiantis // Informatica. –2007. – №31. – P. 249–268.

20. Povkhan I. F. Features of synthesis of generalized features in the construction of recognition systems using the logical tree method / I. F. Povkhan // Information technologies and computer modeling ITKM-2019 : materials of the international scientific and practical conference, Ivano- Frankivsk, May 20–25, 2019. – Ivano-Frankivsk, 2019. – P. 169–174.

21. Vasilenko Y. A. The importance of discrete signs / Y. A. Vasilenko, F. G. Vashuk, I. F. Povkhan // XX International Conference Promising ways and directions of improving the educational system, Uzhgorod, November 16–19, 2010. – Uzhgorod, 2010. – Vol. 21, № 1. – P. 217–222.

22. Deng H. Bias of importance measures for multi-valued attributes and solutions / H. Deng, G. Runger, E. Tuv // Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN), Espoo, Finland, Jun 14–Jun 17, 2011. – Espoo, 2011. – P. 293–300.

23. Kamiński B. A framework for sensitivity analysis of decision trees / B. Kamiński, M. Jakubczyk, P. Szufel // Central European Journal of Operations Research. – 2017. – Vol. 26 (1). – P. 135–159.

24. Dietterich T. G. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization / T. G. Dietterich // Machine learning. – 2000. – Vol. 40, № 2. – P. 139–157.

25. Povhan I. Generation of elementary signs in the general scheme of the recognition system based on the logical tree / I. Povhan // Collection of proceedings «Electronics and information technology». – 2019. – Vol. 12. – P. 20–29.

26. Subbotin S. The dimensionality reduction methods based on computational intelligence in problems of object classification and diagnosis / S. Subbotin, A. Oliinyk // Recent Advances in Systems, Control and Information Technology / eds.: R. Szewczyk, M. Kaliczyńska. – Cham : Springer, 2017. – P. 11–19. – (Advances in Intelligent Systems and Computing, vol. 543).

27. Subbotin S. A. Methods and characteristics of locality-preserving transformations in the problems of computational intelligence / S. A. Subbotin // Radio Electronics, Computer Science, Control. – 2014. – № 1. – P. 120–128.

28. Two-level clustering approach to training data instance selection: a case study for the steel industry / [H. Koskimaki, I. Juutilainen, P. Laurinen, J. Roning] // Neural Networks : International Joint Conference (IJCNN-2008), Hong Kong, 1–8 June 2008 : proceedings. – Los Alamitos : IEEE, 2008. – P. 3044–3049. DOI: 10.1109/ijcnn.2008.4634228

29. Subbotin S. The neuro-fuzzy network synthesis and simplification on precedents in problems of diagnosis and pattern recognition / S. Subbotin // Optical Memory and Neural Networks (Information Optics). – 2013. – Vol. 22, № 2. – P. 97–103. DOI: 10.3103/s1060992x13020082

30. Subbotin S. A. Methods of sampling based on exhaustive and evolutionary search / S. A. Subbotin // Automatic Control and Computer Sciences. – 2013. – Vol. 47, № 3. – P. 113–121. DOI: 10.3103/s0146411613030073

31. De Mántaras R. L. A distance-based attribute selection measure for decision tree induction / De Mántaras R. L. // Machine learning. – 1991. – Vol. 6, № 1. – P. 81–92.

32. Alpaydin E. Introduction to Machine Learning / E. Alpaydin. – London : The MIT Press, 2010. – 400 p.

33. Painsky A. Cross-validated variable selection in tree-based methods improves predictive performance / A. Painsky, S. Rosset // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017. – Vol. 39, № 11. – P. 2142–2153. DOI:10.1109/tpami.2016.2636831

34. Miyakawa M. Criteria for selecting a variable in the construction of efficient decision trees / M. Miyakawa // IEEE Transactions on Computers. – 1989. – Vol. 38, № 1. – P. 130–141.