

ПОСТРОЕНИЕ ДЕРЕВЬЕВ РЕШЕНИЙ ДЛЯ СЛУЧАЯ МАЛОИНФОРМАТИВНЫХ ПРИЗНАКОВ

Субботин С. А. – д-р техн. наук, профессор, заведующий кафедрой программных средств Запорожского национального технического университета, Запорожье, Украина.

АННОТАЦИЯ

Актуальность. Рассмотрена задача автоматизации построения деревьев решений. Объектом исследования являются деревья решений. Предметом исследования являются методы построения деревьев решений.

Цель. Цель работы – создание метода построения моделей на основе деревьев решений для выборок данных, характеризующихся наборами индивидуально малоинформативных признаков.

Метод. Предложен метод построения деревьев решений, который для заданной выборки определяет индивидуальную информативность признаков по отношению к выходному признаку, а также оценивает взаимосвязь входных признаков между собой как их индивидуальную информативность попарно по отношению друг к другу, на этапе формирования очередного узла в качестве признака-кандидата выбирает признак, дающий наилучшее разбиение из всего множества признаков, после чего последовательно ищет среди всех еще неотобранных для данного узла признаков такой признак, который индивидуально наиболее тесно связан с отобранным кандидатом, после чего для совокупности отобранных признаков, перебирая поочередно из заданного набора доступные преобразования, определяет качество разбиения для каждого преобразования, выбирает наилучшее преобразование и добавляет его в узел. Метод стремится при формировании очередного узла выделить группу наиболее тесно взаимосвязанных признаков, преобразование которой в скалярное значение обеспечит наилучшее разбиение подвыборки экземпляров, попавших в данный узел. Это позволяет сократить размер модели и ветвистость дерева, ускорить вычисления при распознавании экземпляров на основе модели, а также повысить обобщающие свойства модели и ее интерпретируемость. Предложенный метод позволяет использовать построенное дерево решений для оценки значимости признаков.

Результаты. Разработанный метод реализован программно и исследован при решении задачи классификации сигналов, представленных набором отсчетов, обладающих индивидуально низкой информативностью.

Выводы. Проведенные эксперименты подтвердили работоспособность предложенного математического обеспечения и позволяют рекомендовать его для использования на практике при решении задач диагностирования и автоматической классификации по признакам. Перспективы дальнейших исследований могут заключаться в создании параллельных методов построения деревьев решений на основе предложенного метода, оптимизации его программных реализаций, а также экспериментального исследования предложенного метода на более широком комплексе практических задач.

КЛЮЧЕВЫЕ СЛОВА: дерево решений, распознавание образов, классификация, признак, информативность.

АББРЕВИАТУРЫ

ДР – дерево решений.

НОМЕНКЛАТУРА

H – набор доступных преобразований;

Λ – множество отобранных признаков для текущего узла;

M – множество неотобранных и нерассмотренных признаков для текущего узла;

Φ – текущее подмножество экземпляров (подвыборка);

ε – граничное приемлемое значение ошибки модели;

$\eta_b(x^s)$ – b -е преобразование;

v – характеристика памяти ДР;

τ – характеристика скорости вычислений ДР;

B – число преобразований в наборе H ;

E_k – ошибка k -го класса классов;

E_{tr} – ошибка модели для обучающих данных;

E_{test} – ошибка модели для тестовых данных;

F – критерий качества модели на основе ДР;

f – структура ДР;

I_j – информативность j -го входного признака;

I^Q – интегральный показатель качества модели в виде ДР;

$I_{i,j}$ – взаимосвязь входных i -го и j -го признаков;

I_{gen} – показатель обобщения моделью обучающей выборки данных;

I_j^{**} – частичная косвенная информативность исходного признака;

I_j^* – частичная явная информативность исходного признака;

I_j – индивидуальная информативность входного признака x_j по отношению к выходному признаку y ;

$J_{\eta_b(\Lambda)}(\Phi)$ – качество разбиения для b -го преобразования;

$J_j(\Phi)$ – критерий разбиения по признаку x_j для экземпляров текущего подмножества Φ в узле;

k – номер класса;

K – число классов в выборке;

N – число входных признаков;

N^l – число используемых в модели признаков;

$N_{l/U}$ – доля листовых узлов в общем числе узлов ДР;

$N_{f/U}$ – доля узлов, содержащих преобразования двух и более признаков в общем числе узлов;

$N_{w/L}$ – среднее число настраиваемых параметров на связь ДР;

$N_{S/l}$ – среднее число экземпляров обучающей выборки на листовую узел ДР;

N_w^L – число настраиваемых параметров связей ДР;

N_w^U – число настраиваемых параметров узлов ДР;

N_U – число узлов в ДР;

N_f^U – число узлов ДР, содержащих преобразования двух и более признаков;

N_w^f – число настраиваемых параметров преобразований модели;

N_i – доля числа использованных моделью признаков в исходном числе входных признаков;

$N_{w/U}$ – среднее число настраиваемых параметров на узел ДР;

$N_{L/U}$ – среднее число связей на узел ДР;

N_l^U – число листовых узлов ДР;

N_L – число связей между узлами;

S – число экземпляров в выборке;

U – множество узлов ДР;

w – множество параметров ДР;

x – множество экземпляров обучающей выборки;

x^s – s -й экземпляр выборки;

x_j – входной j -й признак;

x_j^s – значение j -го входного признака s -го экземпляра выборки;

x_{j^*} – признак, который характеризуется наилучшим значением критерия разбиения текущего подмножества $J_j(\Phi)$ среди всего набора признаков для данного узла;

y – выходной признак;

y^s – значение выходного признака для s -го экземпляра выборки;

Z – критерий останова при построении ДР.

ВВЕДЕНИЕ

Задачи автоматизации поддержки принятия решений возникают во многих практических приложениях. Особую актуальность такие задачи приобретают в случае, когда экспертные знания в предметной области отсутствуют или являются недостаточными, но имеются точечные наблюдения за состоянием объекта или процесса, в отношении которого ставится задача автоматизации принятия решений.

Для такого класса задач весьма эффективным инструментом являются ДР [1–14], представляющие собой получаемые индуктивно модели, обучаемые на основе прецедентов – наблюдений за состоянием моделируемого объекта или процесса.

Объектом исследования являются ДР.

ДР как вычислительная парадигма (класс моделей и методов их построения) обладают такими достоинствами, как простота интерпретации результатов модели, наглядность восприятия и визуализируемость модели в виде дерева, возможность одновременной работы как с категориальными, так и с количественными признаками, отсутствие требований по подготовке данных (нормализации признаков, удаления пропусков), универсальность в плане решения задач классификации и регрессии, возможность работы с пропусками в данных (пустыми значениями признаков) и заполнения пропусков в данных наиболее вероятными значениями, хорошая производительность в процессе классификации по уже построенному дереву (так как метод поиска в дереве весьма эффективен даже для больших наборов данных), возможность оценить надёжность модели при помощи статистических тестов, высокая скорость обучения и работы модели, малое число параметров метода построения модели, возможность определить худшие, лучшие и ожидаемые значения для разных сценариев, способность извлекать знания из данных, учитывать в модели имеющиеся экспертные знания, возможность комбинироваться с другими методами принятия решений [2, 4, 11, 12].

Однако ДР характеризуются рядом недостатков, к которым можно отнести нестабильность процесса построения и высокую чувствительность моделей к шумам и изменениям во входных данных (небольшие изменения в наборе данных могут приводить к построению совершенно другого ДР, что связано с иерархичностью ДР: изменения в узле на верхнем уровне ведут к изменениям во всем дереве ниже), сложность контроля размера ДР (размер ДР является критическим фактором, определяющим качество решения задачи: при использовании простых критериев останова дерева часто растут или очень короткими, или очень большими), неадекватность разделения на классы в сложных случаях (разбиение в узлах происходит по значению одного атрибута параллельно осям признаков, формируя в пространстве признаков прямоугольные кластеры, форма которых может не соответствовать форме реальных кластеров), возможность переобучения из-за создания в процессе построения ДР слишком сложной конструкции, которая недостаточно полно представляет данные, большое число ошибок при использовании излишне детализированных деревьев вследствие переобучения, высокую сложность решения задачи построения оптимального (минимального по размеру и способного без ошибок классифицировать выборку) ДР (NP-полная задача), неэффективность модели для представления задач

типа XOR, чётности или мультиплексарности, пригодность модели только для интерполяции, но не экстраполяции, относительную неточность расчетов (многие другие методы работают лучше с аналогичными данными), сложность вычислений для связанных между собой входных и / или выходных признаков [1, 3, 6, 10, 14].

Предметом исследования являются методы построения ДР.

Известны различные методы построения ДР [1–14], общим свойством которых является ориентация на обработку наборов признаков, содержащих индивидуально информативные признаки. Использование известных методов для обработки сигналов, содержащих большое число индивидуально малоинформативных признаков приведет к созданию деревьев огромного размера, требующих больших затрат памяти на хранение модели и существенных затрат вычислительных ресурсов на расчеты по модели. Поэтому необходимо разработать математическое обеспечение для построения ДР на основе индивидуально малоинформативных признаков.

Цель работы – создание метода построения моделей на основе ДР для выборок данных, характеризующихся наборами индивидуально малоинформативных признаков.

1 ПОСТАНОВКА ПРОБЛЕМЫ

Пусть задана обучающая выборка данных $\langle x, y \rangle$, $x = \{x^s\}$, $y = \{y^s\}$, $y^s \in \{k\}$, $x^s = \{x_j^s\}$, $j = 1, 2, \dots, N$, $s = 1, 2, \dots, S$, $k = 1, 2, \dots, K$. Тогда задача построения ДР на основе выборки $\langle x, y \rangle$ будет состоять в том, чтобы сформировать структуру ДР f и настроить значения его параметров w : $y^s = f(w, x^s)$, $F(f(w, x), y) \rightarrow opt$.

2 ЛИТЕРАТУРНЫЙ ОБЗОР

ДР – модель принятия решений, представленная в виде дерева, узлы которого содержат проверки соответствия значений признаков распознаваемого экземпляра, а связи определяют переходы между узлами для получения решения для распознаваемого экземпляра.

Обобщенно процесс построения ДР можно представить в следующей форме [2, 12].

1. Инициализация. Задать обучающую выборку $\langle x, y \rangle$. Определить критерий останова: достижение максимально допустимой высоты ДР Z , либо получение в каждом листе ДР не более Z экземпляров. В качестве текущего подмножества экземпляров (подвыборки) Φ для корневого узла задать всю исходную выборку $\Phi = \langle x, y \rangle$.

2. Для текущего подмножества экземпляров Φ сформировать текущий узел ДР [15–18]:

– сформировать текущий узел как листовой (у которого нет потомков) и занести в него все текущее подмножество экземпляров;

– проверить критерий останова и при его удовлетворении завершить формирование ДР;

– для каждого признака x_j найти значение критерия разбиения по признаку x_j для экземпляров текущего подмножества в данном узле $J_j(\Phi)$, такой, что чем больше будет его значение, тем лучше будет разбиение [19–26];

– выбрать из набора признаков такой признак x_j^* , который характеризуется наилучшим значением критерия разбиения текущего подмножества $J_j(\Phi)$ среди всего набора признаков для данного узла:

$$j^* = \arg \max_j \{J_j(\Phi)\};$$

– определить текущий узел ДР как внутренний (у которого есть потомки), в который занести признак x_j^* и значение порога разбиения экземпляров по данному признаку;

– выделить подмножества экземпляров, на которые разбивается исходное для данного узла подмножество экземпляров выборки;

– для каждого подмножества рекурсивно выполнить этап формирования текущего узла, сформировав дочерние узлы, после чего установить связи от текущего узла к дочерним узлам.

3. При необходимости для слишком ветвистого ДР выполнить процедуру обрезки ветвей (pruning) [11].

4. Вернуть полученное ДР как результат.

Известные методы построения ДР [1–14] отличаются в основном используемыми процедурами выделения подвыборок, используемыми критериями оценки ценности признаков для помещения их в узел, а также процедурами обрезки ветвей. Однако, как отмечено выше, они предполагают, что набор признаков содержит индивидуально информативные признаки, позволяющие обеспечить эффективное разделение экземпляров на классы.

В случае, когда признаки индивидуально малоинформативны, известные методы будут производить очень глубокие и ветвистые деревья, которые могут не обеспечить требуемую точность, будут обладать низким уровнем обобщения данных, а также будут требовать больших ресурсов памяти для хранения и вычислительно окажутся затратными.

Во многих задачах доступные для измерения признаки являются индивидуально малоинформативными, однако в совокупности обладают достаточной групповой информативностью для принятия решений, а также разделения классов в узлах ДР.

Одним из традиционно применяемых подходов к построению моделей в таких задачах является создание искусственных (конструктивных) признаков на основе взаимосвязанных первичных. Эта задача решается либо пользователем, являющимся экспертом в предметной области, либо путем анализа и обработки сигналов с последующей заменой фрагмента сигнала на его параметры, обобщенно характеризующие важнейшие для задачи свойства [27]. Такой подход, одна-

ко, не позволяет полностью автоматизировать процесс построения ДР и фактически не связан прямо с процессом построения ДР и может рассматриваться лишь как предобработка данных для синтеза ДР.

Поэтому актуальной является задача разработки метода построения ДР для случая индивидуально малоинформативных признаков с автоматическим формированием конструктивных признаков в узлах ДР.

3 МАТЕРИАЛЫ И МЕТОДЫ

Для решения задачи построения ДР по выборке, характеризующейся индивидуально малоинформативными признаками, предлагается перед построением ДР для заданной обучающей выборки определить показатели, характеризующие индивидуальную информативность признаков по отношению к выходному признаку [28, 29], а также определить взаимосвязь входных признаков между собой как их индивидуальную информативность попарно по отношению друг к другу.

На этапе формирования очередного узла в качестве признака-кандидата выбирать признак, дающий наилучшее разбиение из всего множества признаков, после чего последовательно искать среди всех еще неотобранных для данного узла признаков такой признак, который индивидуально наиболее тесно связан с отобранным кандидатом, после чего для совокупности отобранных признаков, перебирая поочередно из заданного набора доступные преобразования, определить качество разбиения для каждого преобразования [30]. Если будет найдено такое преобразование, которое имеет лучшее качество разбиения при добавлении дополнительного признака, чем ранее отобранная для соответствующего узла комбинация признаков, то добавить очередной признак и преобразование в комбинацию. Если же очередной признак не будет улучшать разбиение, то прекратить добавление признаков в узел.

Набор преобразований целесообразно задавать с учетом специфики решаемой задачи. Желательно обеспечить такие преобразования, которые являются достаточно простыми с вычислительной точки зрения и содержательно могут быть интерпретированы, что обеспечит сохранение интерпретируемости модели.

В качестве таких преобразований можно рекомендовать использовать:

– сумму значений признаков группы:

$$\eta_b(x^s) = \sum_{j=1}^N \{x_j^s \mid x_j \in \Lambda\};$$

– произведение значений признаков группы:

$$\eta_b(x^s) = \prod_{j=1}^N \{x_j^s \mid x_j \in \Lambda\};$$

– максимум значений признаков группы:

$$\eta_b(x^s) = \max_{j=1,2,\dots,N} \{x_j^s \mid x_j \in \Lambda\};$$

– минимум значений признаков группы:

$$\eta_b(x^s) = \min_{j=1,2,\dots,N} \{x_j^s \mid x_j \in \Lambda\};$$

– среднее арифметическое значений признаков группы:

$$\eta_b(x^s) = \frac{1}{|\Lambda|} \sum_{j=1}^N \{x_j^s \mid x_j \in \Lambda\};$$

– Эвклидово расстояние в пространстве признаков группы от распознаваемого экземпляра до центра масс экземпляров, попавших в разбиаемый узел:

$$\eta_b(x^s) = \sqrt{\sum_{j=1}^N \left\{ \left(C_j^u - x_j^s \right)^2 \mid x_j \in \Lambda \right\}},$$

где

$$C_j^u = \frac{1}{|\Phi|} \sum_{s=1}^S \{x_j^s \mid x^s \in \Phi\},$$

$x_j \in \Lambda, j = 1, 2, \dots, N.$

– локально чувствительный хэш значений признаков группы [30].

Формально с учетом изложенных идей метод построения ДР можно представить следующим образом.

1. Инициализация. Задать обучающую выборку $\langle x, y \rangle$. Определить критерий останова: достижение максимально допустимой высоты ДР Z либо получение в каждом листе ДР не более Z экземпляров. В качестве текущего подмножества экземпляров (подвыборки) Φ для корневого узла задать всю исходную выборку $\Phi = \langle x, y \rangle$. Задать набор доступных преобразований для объединения исходных признаков в конструируемые $H = \{\eta_b\}$, $b = 1, 2, \dots, B$, где η_b – b -е преобразование из набора, B – число преобразований в наборе.

2. Определение априорной информации об обучающих данных. Оценить индивидуальную информативность I_j каждого входного признака $x_j, j = 1, 2, \dots, N$, по отношению к выходному признаку y . Оценить взаимосвязь входных i -го и j -го признаков как индивидуальную информативность i -го входного признака по отношению к j -му входному признаку $I_{i,j}$.

3. Для текущего подмножества экземпляров Φ сформировать текущий узел ДР:

– сформировать текущий узел как листовой (у которого нет потомков) и занести в него все текущее подмножество экземпляров;

– проверить критерий останова и при его удовлетворении завершить формирование ДР;

– для каждого признака x_j найти значение критерия разбиения по признаку x_j для экземпляров текущего подмножества в данном узле $J_j(\Phi)$, такой, что чем больше будет его значение, тем лучше будет разбиение.

– выбрать из набора признаков такой признак x_{j^*} , который характеризуется наилучшим значением критерия разбиения текущего подмножества $J_j(\Phi)$ среди всего набора признаков для данного узла:

$$j^* = \arg \max_j \{J_j(\Phi)\};$$

– если найдено более одного признака, обладающего наилучшим значением критерия разбиения ($|j^*| > 1$), то из множества таких признаков j^* выбрать тот, который индивидуально является наиболее информативным по отношению к выходному признаку:

$$j^* = \arg \max_p \{I_p \mid p \in \arg \max_j \{J_j(\Phi)\}\},$$

принять: $\Lambda = \{x_{j^*}\}$ – множество отобранных признаков для текущего узла, $M = \{x_i\} \setminus x_{j^*}$ – множество неотобранных и нерассмотренных для текущего узла признаков, $i = 1, 2, \dots, N$.

4. Группировка признаков в текущем узле u :

– если отсутствуют еще неотобранные и нерассмотренные для текущего узла признаки ($|M|=0$), то перейти к следующему этапу;

– среди всех еще неотобранных и нерассмотренных для текущего узла признаков определить такой признак, который индивидуально наиболее тесно связан с отобранным признаком кандидатом j^* :

$$\Lambda = \Lambda \cup \{x_p \mid p = \arg \max_i \{I_{i,j^*} \mid x_i \notin \Lambda, x_i \in M\}\},$$

после чего установить: $M = M \setminus x_p$;

– для совокупности отобранных признаков Λ , перебирая поочередно из заданного набора N доступные преобразования, определить качество разбиения для каждого преобразования: $J_{\eta_b(\Lambda)}(\Phi)$, $b = 1, 2, \dots, B$.

– если очередной j^* -й признак не будет улучшать разбиение ($\forall b = 1, 2, \dots, B: J_{\eta_b(\Lambda)}(\Phi) \leq J_{\eta_b(\Lambda \setminus x_{j^*})}(\Phi)$), то прекратить добавление признаков в узел, установив $\Lambda = \Lambda \setminus x_{j^*}$ и перейдя к следующему этапу; в противном случае, если будет найдено такое преобразование, которое имеет лучшее качество разбиения при добавлении дополнительного признака, чем ранее отобранная для соответствующего узла комбинация

признаков ($\exists b = 1, 2, \dots, B: J_{\eta_b(\Lambda)}(\Phi) > J_{\eta_b(\Lambda \setminus x_{j^*})}(\Phi)$), то

добавить очередной признак и преобразование в комбинацию, повторив выполнение этапа группировки признаков в узле.

5. Этап разбиения текущего узла и формирования ветвей:

– определить текущий узел ДР как внутренний (у которого есть потомки), в который занести признак x_{j^*} и значение порога разбиения экземпляров по данному признаку;

– выделить подмножества экземпляров, на которые разбивается исходное для данного узла подмножество экземпляров выборки;

– для каждого подмножества рекурсивно выполнить этапы 3–5, сформировав дочерние узлы, после чего установить связи от текущего узла к дочерним узлам.

6. При необходимости для слишком ветвистого ДР выполнить процедуру обрезки ветвей (pruning).

7. Вернуть полученное ДР как результат.

Предложенный метод для класса задач с индивидуально малоинформативными признаками будет стремиться при формировании очередного узла выделить группу наиболее тесно взаимосвязанных признаков, преобразование которой в скалярное значение обеспечит наилучшее разбиение подвыборки экземпляров, попавших в данный узел. Это позволит, во-первых, сократить размер модели и ветвистость ДР, во-вторых, ускорит вычисления при распознавании экземпляров на основе модели, а в-третьих, повысит обобщающие свойства модели и ее интерпретируемость. Вместе с этим предложенный метод может оказаться более вычислительно затратным на этапе построения модели из-за необходимости большего числа переборов преобразований и сочетаний признаков для групп в каждом узле.

ДР, построенное с помощью предложенного метода, возможно также использовать для оценки значимости признаков.

1. Задать ДР и исходную выборку.

2. Определить признаки, использованные явно в узлах модели.

3. Определить признаки, использованные неявно для определения хэшей, использованных в узлах модели.

3. Все признаки, которые не входят ни в , считать неинформативными, приняв для них $I_j = 0$.

4. Определить частичную явную информативность исходного признака, использованного отдельно (без преобразований) во внутреннем узле, как долю наименьшего числа экземпляров, попавших в дочерний узел $|\Phi|$, в исходной выборке:

$$I_j^* = \frac{|\Phi|}{S}.$$

5. Определить частичную косвенную информативность исходного признака, использованного в преобразовании во внутреннем узле, как долю наименьшего числа экземпляров, попавших в дочерний узел, в исходной выборке, деленную на $|\Lambda|$ – число признаков, входящих в группу преобразования для соответствующего узла:

$$I_j^{**} = \frac{|\Phi|}{S|\Lambda|}.$$

6. Определить совокупную информативность признака как максимальную из его частичных явной и косвенной информативностей для всего множества узлов ДР U :

$$I_j = \max(\max_U \{I_j^*\}, \max_U \{I_j^{**}\}).$$

7. Нормировать значения информативностей признаков так, чтобы их сумма равнялась единице:

$$I_j = \frac{I_j}{\sum_{i=1}^N I_i}.$$

4 ЭКСПЕРИМЕНТЫ

Поскольку для одной и той же обучающей выборки данных с помощью различных методов и при разных значениях их параметров возможно получить различные модели в виде ДР на практике возникает задача сравнения имеющихся моделей и выбора наилучшей модели из них.

Для сравнения моделей и методов построения ДР целесообразно определить показатели, характеризующие их наиболее важные свойства, с последующим определением на их основе интегрального показателя качества – критерия сравнения моделей на основе ДР.

К элементарным характеристикам ДР (с указанием их граничных значений) относятся:

- число узлов в дереве $N_U \geq K+1$;
- число связей между узлами $N_L > N_U$;
- число настраиваемых параметров узлов ДР $N_w^U > 0$;
- число настраиваемых параметров связей ДР $N_w^L \geq 0$;
- число узлов ДР, содержащих преобразования двух и более признаков, $N_f^U \geq 0$;
- число листовых узлов ДР $N_l^U \geq K$;
- число используемых в модели признаков $N', N \geq N' \geq 1$;
- ошибка модели для обучающих данных $E_{tr} \geq 0$;
- ошибка модели для тестовых данных $E_{test} \geq 0$;

– ошибки для каждого из классов $E_k \geq 0, k = 1, 2, \dots, K$.

Определим относительные характеристики ДР на основе его элементарных характеристик:

– среднее число связей на узел ДР:

$$N_{L/U} = \frac{N_L}{N_U} \geq 0;$$

– среднее число настраиваемых параметров на узел ДР:

$$N_{w/U} = \frac{N_w^U}{N_U} \geq 0;$$

– среднее число настраиваемых параметров на связь ДР:

$$N_{w/L} = \frac{N_w^L}{N_L} \geq 0;$$

– долю числа использованных моделью признаков в исходном числе входных признаков:

$$N_l = \frac{N'}{N} > 0;$$

– долю узлов, содержащих преобразования двух и более признаков в общем числе узлов:

$$N_{f/U} = \frac{N_f^U}{N_U} \geq 0;$$

– долю листовых узлов в общем числе узлов ДР:

$$N_{l/U} = \frac{N_l^U}{N_U} \geq K;$$

– среднее число экземпляров обучающей выборки на листовой узел ДР:

$$N_{S/l} = \frac{S}{N_l^U} > 0.$$

Показатель обобщения моделью обучающей выборки данных:

$$I_{gen} = \frac{NS}{N_w^f + N_w^U + N_U + 2N_L} > 0,$$

где N_w^f – число настраиваемых параметров преобразований модели.

Важнейшими параметрами модели на основе ДР, значения которых необходимо минимизировать, являются ошибки модели для обучающих данных E_{tr} , для тестовых данных E_{test} и ошибки для каждого из классов E_k .

Среди параметров ДР при приемлемой ошибке целесообразно уменьшать показатели структурной сложности (число используемых в модели признаков N' , долю числа использованных моделью признаков в исходном числе входных признаков N_l , число узлов в дереве N_U и число связей между узлами N_L), показатели параметрической сложности (число настраиваемых параметров узлов ДР N_w^U и число настраиваемых параметров связей ДР N_w^L , среднее число настраиваемых параметров на узел ДР $N_{w/U}$, среднее число настраиваемых параметров на связь ДР $N_{w/L}$, среднее число связей на узел: $N_{L/U}$), характеристики памяти v и скорости вычислений τ .

Увеличивать целесообразно долю листовых узлов в общем числе узлов ДР $N_{l/U}$ и число листовых узлов ДР N_l^U (это будет сокращать промежуточные уровни принятия решений), число узлов ДР, содержащих преобразования двух и более признаков N_f^U и долю узлов, содержащих преобразования двух и более признаков в общем числе узлов, $N_{f/U}$ (это будет обеспечивать повышение обобщения данных и сокращать ветвистость ДР), а также показатель обобщения I_{gen} , среднее число экземпляров обучающей выборки на листовой узел ДР $N_{S/l}$.

Интегральный показатель качества модели в виде ДР с учетом изложенных выше соображений, подставив определения вторичных показателей через первичные, приведя подобные, исключив дуближж показателей, конструктивно можно определить как:

$$I^Q = \begin{cases} (N_f^U + N_l^U) e^{-\frac{|E_{tr} E_{test} - \varepsilon^2|}{SS_{test}}} \\ N' N_U N_L (N_w^L + N_w^U + N_w^f), \frac{E_{tr}}{S} \leq \varepsilon; \\ 0, \frac{E_{tr}}{S} > \varepsilon. \end{cases}$$

Предложенный показатель для средней ошибки, большей заданного максимально допустимого уровня, будет равен нулю. В случае, если средняя ошибка является приемлемой, показатель будет принимать значения от нуля до суммы числа листовых узлов и внутренних узлов, содержащих преобразования двух и более признаков. Чем меньше будет значение показателя, тем хуже модель по совокупности свойств. Чем

больше будет значение показателя, тем лучше будет модель по совокупности свойств.

Разработанный метод построения ДР и предложенный комплекс показателей качества ДР были программно реализованы при решении задач классификации сигналов: задачи диагностики лопаток газотурбинных авиадвигателей [31] и задачи классификации растительных объектов по результатам дистанционного зондирования [32].

Общим свойством решавшихся задач являлось то, что входные данные представляли собой наборы отсчетов сигнала (или спектров сигнала), которые индивидуально являлись малоинформативными.

Разработанное программное обеспечение использовалось для проведения экспериментов по исследованию работы предложенного метода при решении практических задач классификации по индивидуально малоинформативным признакам.

5 РЕЗУЛЬТАТЫ

Фрагмент результатов проведенных экспериментов представлен в табл. 1.

В обеих задачах построенные модели ДР обладали приемлемым уровнем точности, заданным с учетом требований задачи.

6 ОБСУЖДЕНИЕ

Разработанный комплекс показателей качества моделей на основе ДР позволяет количественно выразить важнейшие свойства деревьев решений и может быть использован в процессе построения моделей как критерий оптимизации, а также для сравнения и выбора итоговой модели из множества моделей, построенных на основе различных методов синтеза ДР.

Как видно из табл. 1, предложенный метод построения моделей на основе ДР по сравнению с классическим подходом [1] позволяет синтезировать модели, обладающие приемлемой точностью, но более высоким качеством, что объясняется более высоким уровнем обобщения данных, что, в свою очередь, обеспечивается минимизацией числа узлов-проверок и автоматическим формированием преобразований исходных признаков в узлах.

ВЫВОДЫ

Решена задача автоматизации построения ДР по набору прецедентов.

Научная новизна полученных результатов состоит в том, что впервые предложен метод построения ДР, который для заданной выборки определяет индивидуальную информативность признаков по отношению к выходному признаку, а также оценивает взаимосвязь входных признаков между собой как их индивидуальную информатив-

Таблица 1 – Характеристики задач и моделей в виде ДР

Задача	Характеристики исходной выборки			I^Q полученной модели ДР	
	K	N	S	Метод [2]	Предложенный метод
Диагностирование лопаток газотурбинных авиадвигателей [31]	2	1000	32	0,000521669	0,000561664
Классификация растительных объектов на культурные и сорные по результатам дистанционного зондирования [32]	2	55	248	0,000044837	0,000055713

ность попарно по отношению друг ко другу, на этапе формирования очередного узла в качестве признака-кандидата выбирает признак, дающий наилучшее разбиение из всего множества признаков, после чего последовательно ищет среди всех еще неотобранных для данного узла признаков такой признак, который индивидуально наиболее тесно связан с отобранным кандидатом, после чего для совокупности отобранных признаков, перебирая поочередно из заданного набора доступные преобразования, определяет качество разбиения для каждого преобразования, выбирает наилучшее преобразование и добавляет его в узел. Метод стремится при формировании очередного узла выделить группу наиболее тесно взаимосвязанных признаков, преобразование которой в скалярное значение обеспечит наилучшее разбиение подвыборки экземпляров, попавших в данный узел. Это позволяет сократить размер модели и ветвистость ДР, ускорить вычисления при распознавании экземпляров на основе модели, а также повысить обобщающие свойства модели и ее интерпретируемость. Предложенный метод позволяет использовать построенное ДР для оценки значимости признаков.

Предложен комплекс показателей, позволяющих количественно характеризовать свойств моделей на основе ДР, который может использоваться как для сравнения и выбора наилучшей модели, так и в процессе построения моделей для определения критерия оптимальности моделей.

Практическая ценность полученных результатов состоит в том, что разработанный метод реализован программно и исследован при решении задачи классификации сигналов, представленного набором отсчетов, обладающих индивидуально низкой информативностью.

Проведенные эксперименты подтвердили работоспособность предложенного математического обеспечения и позволяют рекомендовать его для использования на практике при решении задач диагностирования и автоматической классификации по признакам.

Перспективы дальнейших исследований могут заключаться в создании параллельных методов построения ДР на основе предложенного метода, оптимизации его программных реализаций, а также экспериментальном исследовании предложенного метода на более широком комплексе практических задач.

БЛАГОДАРНОСТИ

Работа выполнена в рамках госбюджетной научно-исследовательской темы «Интеллектуальные методы и программные средства диагностирования и неразрушающего контроля качества техники военного и гражданского назначения» Запорожского национального технического университета при частичной поддержке международного проекта «Innovative Multidisciplinary Curriculum in Artificial Implants for Bio-Engineering BSc/MSc Degrees» (BIOART, регистрационный номер 586114-EPP-1-2017-1-ES-EPPKA2-SVHE-JP) программы «Эразмус+» Европейского Союза.

ЛИТЕРАТУРА / ЛІТЕРАТУРА

1. 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. – V. 19, № 11. – P. 1300–1305.
2. Classification and regression trees / [L. L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone]. – Boca Raton : Chapman and Hall/CRC, 1984. – 368 p.

3. 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/tr-bias.pdf>
4. Geurts P. Supervised learning with decision tree-based methods in computational and systems biology / P. Geurts, A. Irthum, L. Wehenkel // Molecular Biosystems. – 2009. – V. 5, № 12. – P. 1593–1605.
5. Heath D. Induction of oblique decision trees [Electronic resource] / D. Heath, S. Kasif, S. Salzberg. – Baltimore : Johns Hopkins University, 1993. – 6 p. – Access mode: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.9208&rep=rep1&type=pdf>
6. Hothorn T. Unbiased recursive partitioning: A conditional inference framework / T. Hothorn, K. Hornik, A. Zeileis // Journal of Computational and Graphical Statistics. – 2006. – V. 15, № 3. – P. 651–674.
7. Hyafil L. Constructing optimal binary decision trees is np-complete / L. Hyafil, R. L. Rivest // Information Processing Letters. – 1976. – V. 5, № 1. – P. 15–17.
8. Kim H. Classification trees with unbiased multiway splits / H. Kim, W.-Y. Loh // Journal of the American Statistical Association. – 2001. – V. 96, № 454. – P. 589–604.
9. Kufirin R. Decision trees on parallel processors / R. Kufirin // Machine Intelligence and Pattern Recognition. – 1997. – V. 20. – P. 279–306.
10. Kwok S. W. Multiple decision trees / S. W. Kwok, C. Carter // Fourth Annual Conference on Uncertainty in Artificial Intelligence (UAI '88), 10–12 July 1988, Minneapolis : proceedings. – Amsterdam North-Holland Publishing Co., 1990. – P. 327–338.
11. Mingers J. An empirical comparison of pruning methods for decision tree induction / J. Mingers // Machine learning. – 1989. – V. 4, № 2. – P. 227–243.
12. Quinlan J. R. Induction of decision trees / J. R. Quinlan // Machine learning. – 1986. – V. 1, № 1. – P. 81–106.
13. Strobl C. Unbiased split selection for classification trees based on the Gini index / C. Strobl, A. Boulesteix, T. Augustin. – Computational Statistics & Data Analysis. – 2007. – V. 52, № 1. – P. 483–501.
14. Utgoff P. E. Incremental induction of decision trees / P. E. Utgoff // Machine learning, 1989. – V. 4, № 2. – P. 161–186. DOI:10.1023/A:1022699900025
15. Breiman L. Bagging predictors / L. Breiman // Machine Learning. – 1996. – V. 24, № 2. – P. 123–140.
16. Learning ensembles from bites: A scalable and accurate approach / [N. V. Chawla, L. O. Hall, K. W. Bowyer, W. P. Kegelmeyer] // Journal of Machine Learning Results. – 2004. – V. 5. – P. 421–451.
17. 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. – V. 40, № 2. – P. 139–157.
18. Efron B. Bootstrap methods: another look at the jackknife / B. Efron // The Annals of Statistics. – 1979. – V. 7, № 1. – P. 1–26.
19. Permutation importance: a corrected feature importance measure / [A. Altmann, L. Tološi, O. Sander, T. Lengauer] // Bioinformatics. – 2010. – V. 26, № 10. – P. 1340–1347. DOI: 10.1093/bioinformatics/btq134.
20. De Mántaras R. L. A distance-based attribute selection measure for decision tree induction / R. L. De Mántaras // Machine learning. – 1991. – V. 6, № 1. – P. 81–92.
21. Deng H. Bias of importance measures for multi-valued attributes and solutions / H. Deng, G. Runger, E. Tuv // 21st International Conference on Artificial Neural Networks (ICANN), Espoo, 14–17 June 2011 : proceedings. – Berlin : Springer-Verlag, 2011. – V. 2. – P. 293–300.
22. Mingers J. An empirical comparison of selection measures for decision-tree induction / J. Mingers // Machine learning. – 1989. – V. 3, № 4. – P. 319–342.
23. Mitchell T. Machine learning / T. Mitchell. – New York : McGraw-Hill, 1997. – 432 p.
24. 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. – V. 39, № 11. – P. 2142–2153. DOI:10.1109/tpami.2016.2636831
25. Miyakawa M. Criteria for selecting a variable in the construction of efficient decision trees / M. Miyakawa // IEEE Transactions on Computers. – 1989. – V. 38, № 1. – P. 130–141.
26. Tolosi L. Classification with correlated features: unreliability of feature ranking and solutions / L. Tolosi, T. Lengauer // Bioinformatics. – 2011. – V. 27, № 14. – P. 1986–1994. DOI: 10.1093/bioinformatics/btr300
27. Alpaydin E. Introduction to Machine Learning / E. Alpaydin. – London : The MIT Press. 2010. – 400 p.
28. Subbotin S. Entropy based evolutionary search for feature selection / S. Subbotin, An. Oleynik // The experience of designing and application of CAD systems in microelectronics : IX International conference CADSM-2007, Lviv–Polyana, 20–24 February 2007 : proceedings. – Lviv : NU “Lviv Polytechnic”, 2007. – P. 442–443.
29. 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. Kalicyńska. – Cham: Springer, 2017. – P. 11–19. – (Advances in Intelligent Systems and Computing, vol. 543).
30. 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.
31. Прогрессивные технологии моделирования, оптимизации и интеллектуальной автоматизации этапов жизненного цикла авиационных двигателей : монография / А. В. Богуслаев, Ал. А. Олейник, Ан. А. Олейник, Д. В. Павленко, С. А. Субботин ; под ред. Д. В. Павленко, С. А. Субботина. – Запорожье : ОАО «Мотор Сич», 2009. – 468 с.
32. The plant recognition on remote sensing results by the feed-forward neural networks / [V. Dubrovin, S. Subbotin, S. Morshchavka, D. Piza] // International Journal of Smart Engineering System Design. – 2001. – V. 3, № 4. – P. 251–256.

Received 11.02.2019.
Accepted 18.02.2019.

УДК 004.93

ПОБУДОВА ДЕРЕВ РІШЕНЬ ДЛЯ ВИПАДКУ МАЛОІНФОРМАТИВНИХ ОЗНАК

Субботін С. О. – д-р техн. наук, професор, завідувач кафедри програмних засобів Запорізького національного технічного університету, Запоріжжя, Україна.

АНОТАЦІЯ

Актуальність. Розглянуто задачу автоматизації побудови дерев рішень. Об'єктом дослідження є дерева рішень. Предметом дослідження є методи побудови дерев рішень.

Мета. Мета роботи – створення методу побудови моделей на основі дерев рішень для вибірок даних, що характеризуються наборами індивідуально малоінформативних ознак.

Метод. Запропоновано метод побудови дерев рішень, який для заданої вибірки визначає індивідуальну інформативність ознак по відношенню до вихідної ознаки, а також оцінює взаємозв'язок вхідних ознак між собою як їх індивідуальну інформативність попарно по відношенню один до одного, на етапі формування чергового вузла в якості ознаки-кандидата обирає ознаку, що дає найкраще розбиття з усієї множини ознак, після чого послідовно шукає серед усіх ще невідібраних для даного вузла ознак таку, що індивідуально найтісніше пов'язана з відібраним кандидатом, після чого для сукупності відібраних ознак, перебираючи по черзі з заданого набору доступні перетворення, визначає якість розбиття для кожного перетворення, обирає найкраще перетворення і додає його в вузол. Метод прагне при формуванні чергового вузла виділити групу найбільш тісно взаємопов'язаних ознак, перетворення якої у скалярне значення забезпечить найкращу розбиття підвибірці зразків, які потрапили у даний вузол. Це дозволяє скоротити розмір моделі і гіллястість дерева, прискорити обчислення при розпізнаванні екземплярів на основі моделі, а також підвищити узагальнювальні властивості моделі та її інтерпретабельність. Запропонований метод дозволяє використовувати побудоване дерево рішень для оцінки значимості ознак.

Результати. Розроблений метод реалізований програмно і досліджений під час розв'язання задачі класифікації сигналів, поданих набором відліків, що є індивідуально малоінформативними.

Висновки. Проведені експерименти підтвердили працездатність запропонованого математичного забезпечення і дозволяють рекомендувати його для використання на практиці при вирішенні завдань діагностування та автоматичної клас-класифікації за ознаками. Перспективи подальших досліджень можуть полягати у створенні паралельних методів побудови дерев рішень на основі запропонованого методу, оптимізації його програмних реалізацій, а також експериментальному дослідженні запропонованого методу на більш широкому комплексі практичних завдань.

КЛЮЧОВІ СЛОВА: дерево рішень, розпізнавання образів, класифікація, ознака, інформативність.

UDC 004.93

DECISION TREE CONSTRUCTION FOR THE CASE OF LOW-INFORMATIVE FEATURES

Subbotin S. A. – Dr. Sc., Professor, Head of the Department of Software Tools, Zaporizhzhia National Technical University, Zaporizhzhia, Ukraine.

ABSTRACT

Context. The problem of automating the decision tree construction is addressed. The object of study is a decision tree. The subject of study is the methods of decision tree building.

Objective. The purpose of the work is to create a method for constructing models based on decision trees for data samples that are characterized by sets of individually low-informative features.

Method. A method for decision tree constructing is proposed, which for a given sample determines the individual informativeness of features relatively to the output feature, and also evaluates the relationship of input features with each other as their individual informativity pairwise relatively to each other, at the step of forming the next node the method selects as a candidate feature the feature that gives the best partition in the whole set of features, after which it sequentially searches among all the features that are not selected for this node the one that is individually most closely related with the selected candidate, then for the set of selected features, iterating through the available transformations from a given set, determines the quality of the partition for each transformation, selects the best transformation and adds it to the node. When forming the next node, the method tends to single out a group of the most closely interrelated features, the conversion of which into a scalar value will provide the best partitioning of a subsample of instances hit into this node. This makes possible to reduce the size of the model and the branching of the tree, speed up the calculations in recognizing instances based on the model, as well as improve the generalizing properties of the model and its interpretability. The proposed method allows using the constructed decision tree to assess the feature significance.

Results. The developed method is implemented as software and investigated at signal represented by a set of individually low-informative readings classification problem solving.

Conclusions. The experiments have confirmed the efficiency of the proposed software and allow recommending it for use in practice in solving problems of diagnostics and automatic classification by features. The prospects for further research may consist in the creation of parallel methods for constructing decision trees based on the proposed method, optimization of its software implementations, and also in an experimental study of the proposed method on a wider set of practical problems.

KEYWORDS: decision tree, pattern recognition, classification, feature, informativeness.

REFERENCES

1. Amit Y., Geman D., Wilder K. Joint induction of shape features and tree classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, V. 19, № 11, pp. 1300–1305.
2. 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.
3. 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/tr-bias.pdf>
4. Geurts P., IRRthum A., Wehenkel L. Supervised learning with decision tree-based methods in computational and systems biology, *Molecular Biosystems*, 2009, V. 5, No. 12, pp. 1593–1605.
5. Heath D., Kasif S., Salzberg S. Induction of oblique decision trees [Electronic resource]. Baltimore, Johns Hopkins University, 1993, 6 p. Access mode: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.9208&rep=rep1&type=pdf>
6. Hothorn T., Hornik K., Zeileis A. Unbiased recursive partitioning: A conditional inference framework, *Journal of Computational and Graphical Statistics*, 2006, V. 15, No. 3, pp. 651–674.
7. Hyafil L., Rivest R. L. Constructing optimal binary decision trees is np-complete, *Information Processing Letters*, 1976, V. 5, № 1, pp. 15–17.
8. Kim H., Loh W.-Y. Classification trees with unbiased multi-way splits, *Journal of the American Statistical Association*, 2001, V. 96, No. 454, pp. 589–604.
9. Kufirin R. Decision trees on parallel processors, *Machine Intelligence and Pattern Recognition*, 1997, V. 20, pp. 279–306.
10. Kwok S. W., Carter C. Multiple decision trees, *Fourth Annual Conference on Uncertainty in Artificial Intelligence (UAI '88), 10–12 July 1988, Minneapolis : proceedings*. Amsterdam North-Holland Publishing Co., 1990, pp. 327–338.
11. Mingers J. An empirical comparison of pruning methods for decision tree induction, *Machine learning*, 1989, V. 4, No. 2, pp. 227–243.
12. Quinlan J. R. Induction of decision trees, *Machine learning*, 1986, V. 1, No. 1, pp. 81–106.
13. Strobl C., Boulesteix A., Augustin T. Unbiased split selection for classification trees based on the Gini index, *Computational Statistics & Data Analysis*, 2007, V. 52, No. 1, pp. 483–501.
14. Utgoff P. E. Incremental induction of decision trees, *Machine learning*, 1989, V. 4, No. 2, pp. 161–186. DOI: 10.1023/A:1022699900025
15. Breiman L. Bagging predictors, *Machine Learning*, 1996, V. 24, No. 2, pp. 123–140.
16. Chawla N. V., Hall L. O., Bowyer K. W., W. P. Kegelmeyer Learning ensembles from bites: A scalable and accurate approach, *Journal of Machine Learning Results*, 2004, V. 5, pp. 421–451.
17. Dietterich T. G. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, *Machine learning*, 2000, V. 40, № 2, pp. 139–157.
18. Efron B. Bootstrap methods: another look at the jackknife, *The Annals of Statistics*, 1979, V. 7, № 1, pp. 1–26.
19. Altmann A., Tološi L., Sander O., Lengauer T. Permutation importance: a corrected feature importance measure, *Bioinformatics*, 2010, V. 26, № 10, pp. 1340–1347. DOI:10.1093/bioinformatics/btq134.
20. De Mántaras R. L. A distance-based attribute selection measure for decision tree induction, *Machine learning*, 1991, V. 6, № 1, pp. 81–92.
21. Deng H., Runger G., Tuv E. Bias of importance measures for multi-valued attributes and solutions, *21st International Conference on Artificial Neural Networks (ICANN), Espoo, 14–17 June 2011 : proceedings*. Berlin, Springer-Verlag, 2011, V. 2, pp. 293–300.
22. Mingers J. An empirical comparison of selection measures for decision-tree induction, *Machine learning*, 1989, V. 3, No. 4, pp. 319–342.
23. Mitchell T. Machine learning. New York, McGraw-Hill, 1997, 432 p.
24. Painsky A., Rosset S. Cross-validated variable selection in tree-based methods improves predictive performance, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, V. 39, No. 11, pp. 2142–2153. DOI:10.1109/tpami.2016.2636831
25. Miyakawa M. Criteria for selecting a variable in the construction of efficient decision trees, *IEEE Transactions on Computers*, 1989, V. 38, No. 1, pp. 130–141.
26. Tolosi L., Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions, *Bioinformatics*, 2011, V. 27, No. 14, pp. 1986–1994. DOI:10.1093/bioinformatics/btr300
27. Alpaydin E. Introduction to Machine Learning. London, The MIT Press, 2010, 400 p.
28. Subbotin S., Oleynik An. Entropy based evolutionary search for feature selection, *The experience of designing and application of CAD systems in microelectronics : IX International conference CADSM–2007, Lviv–Polyana, 20–24 February 2007 : proceedings*. Lviv, NU “Lviv Polytechnic”, 2007, pp. 442–443.
29. Subbotin S., Oliinyk A. Eds.: R. Szweczyk, M. Kaliczynska 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).
30. Subbotin S. A. Methods and characteristics of locality-preserving transformations in the problems of computational intelligence, *Radio Electronics, Computer Science, Control*, 2014, No. 1, pp. 120–128.
31. Boguslayev A. V., Oleynik Al. A., Oleynik An. A., Pavlenko D. V., Subbotin S. A.; pod red. Pavlenko D. V., Subbotina S. A. *Progressivnyye tekhnologii modelirovaniya, optimizatsii i intelektual'noy avtomatizatsii etapov zhiznennogo tsikla aviatsonnykh dvigateley : monografiya*. Zaporozh'ye, OAO “Motor Sich”, 2009, 468 p.
32. Dubrovin V., Subbotin S., Morshchavka S., Piza D. The plant recognition on remote sensing results by the feed-forward neural networks, *International Journal of Smart Engineering System Design*, 2001, V. 3, No. 4, pp. 251–256.