УДК 004.93

**Е. А. Гофман[1], А. А. Олейник[2], С. А. Субботин[3]**

[1] Запорожский национальный технический университет, г. Запорожье, Украина, gofman_jenek@rambler.ru;

[2] Запорожский национальный технический университет, г. Запорожье, Украина, olejnikaa@gmail.com;

[3] Запорожский национальный технический университет, г. Запорожье, Украина, subbotin@zntu.edu.ua

# ИНДУКЦИЯ ЛИНГВИСТИЧЕСКИХ ПРАВИЛ С ИСПОЛЬЗОВАНИЕМ ДЕРЕВЬЕВ РЕШЕНИЙ

Рассматривается задача индукции лингвистических правил. Разработан метод идентификации деревьев решений для индукции лингвистических правил. Создано программное обеспечение на основе предложенного метода.

ДЕРЕВО РЕШЕНИЙ, ИДЕНТИФИКАЦИЯ, ИНДУКЦИЯ ПРАВИЛ, ЛИНГВИСТИЧЕСКОЕ ПРАВИЛО, ОБУЧАЮЩАЯ ВЫБОРКА

## Введение

В настоящее время экспертные системы, основанные на лингвистических правилах [1], успешно применяются в различных прикладных областях, таких как медицинская и техническая диагностика, финансовый менеджмент, распознавание образов, геологическая разведка, управление компьютерными сетями, управление технологическими процессами, анализ веб-контента в Интернет и др. Широкое применение таких систем обусловлено в первую очередь тем, что они являются прозрачными и относительно дешёвыми в реализации.

Поскольку базы правил в экспертных системах часто характеризуются большим объёмом, актуальной является задача индукции правил, суть которой заключается в том, что на основе начального набора правил необходимо сформировать новую базу правил меньшего объёма, которая в достаточной мере представляла бы начальную базу правил и была бы менее избыточной.

Существуют различные методы индукции правил [2], однако эти методы при обработке правил анализируют их качество по отдельности, не рассматривая и не учитывая качество всей базы в целом, что приводит к получению неоптимальных баз нечётких правил. Поэтому актуальной является разработка новых методов индукции правил, которые учитывали бы качество всей базы знаний, а не только отдельных правил. Для решения данной задачи предлагается создавать деревья решений [3–5], которые после их построения переводились бы в лингвистические правила. Выбор деревьев решений обосновывается их возможностью выявлять ненаблюдаемые связи внутри рассматриваемых объектов, процессов и систем.

Целью данной работы является разработка метода индукции лингвистических правил с использованием математического аппарата деревьев решений.

Для достижения поставленной цели необходимо решить следующие задачи:

– обзор математического аппарата деревьев решений;

– приведение основных этапов идентификации деревьев решений в соответствие с решаемой задачей;

– создание правил преобразования деревьев решений в лингвистические правила;

– сравнение разработанного подхода с существующими методами индукции лингвистических правил.

## 1. Постановка задачи

Пусть задана база лингвистических правил $RB = \{R^1, R^2 ..., R^{RN}\}$, описывающая объекты обучающей выборки $O = \{O^1, O^2, ..., O^N\}$. Тогда на основе обучающей выборки объектов $O$ требуется сформировать такую базу лингвистических правил $RB^* = \{R^1, R^2, ..., R^{RN^*}\}$, $RN^* << RN$, которая обеспечивала бы приемлемое качество прогнозирования экспертной системы, построенной на основе полученной базы лингвистических правил $RB^*$:

$$Q(RB^*) \geq Q_{threshold} \cdot$$

где $Q(RB^*)$ – точность прогнозирования или классификации по базе правил $RB^*$; $Q_{threshold}$ – минимально допустимая точность прогнозирования или классификации.

## 2. Деревья решений

Деревья решений представляют собой графовые интеллектуальные модели, во внутренних узлах которых расположены функции принятия решений на основе значений входных переменных, а во внешних узлах (терминальных узлах, листьях) содержатся значения выходной переменной, соответствующие условиям внутренних узлов [2, 6, 7].

Благодаря своей древовидной структуре такие модели позволяют наглядно представлять результаты вычислений. Поэтому они хорошо интерпретируются людьми-специалистами в прикладных областях, которые, как правило, не имеют специальной математической подготовки и незнакомы с методами и моделями искусственного интеллекта. Деревья решений позволяют эффективно решать задачи классификации и прогнозирования, обеспечивая при этом высокую точность.

Для применения деревьев решений на практике в целях классификации или прогнозирования значений выходных параметров исследуемых объектов по наборам значений входных характеристик необходимо с помощью данных обучающей выборки сформировать дерево решений таким образом, чтобы оно наилучшим образом описывало исследуемый объект.

Построение деревьев решений связано с извлечением правил из обучающих выборок. Каждый путь от корня дерева к одному из его листьев может быть преобразован к логическому высказыванию – правилу типа «если А, то В», где его антецедент получается путем использования всех условий, представленных во внутренних узлах от корня к выходному листу, а правая часть правила получается из соответствующего листа дерева.

Процесс построения дерева решений, как правило, содержит такие этапы: разрастание, ветвление, вычисление значения выходного параметра для листа, сокращение.

В результате этапа разрастания (увеличения, growing) некоторая вершина заменяется поддеревом, полученным путем ветвления этой вершины. На данном этапе происходит разделение выбранной вершины на несколько новых (в случае дихотомического дерева вершина разбивается на две новых). При этом перебираются все признаки и все возможные варианты ветвления по каждому из признаков. В результате остается вариант разбиения, при котором значение критерия качества разбиения является наилучшим. Если новые вершины являются перспективными для последующего разделения (критерии завершения разрастания не удовлетворены), то выполняется их ветвление. В случае невозможности дальнейшего разделения вершины она становится листом, и для нее выполняется процедура вычисления значения выходного параметра. Если ветвление вершины приводит к ухудшению качества дерева, то вершина также объявляется листом.

Процедура ветвления (разделения, splitting) дерева вызывается рекурсивно при выполнении этапа разрастания. Ветвление подразумевает создание для выбранной вершины заданного количества (для дихотомических деревьев – две) вершин-потомков.

Вычисление значения выходного параметра происходит путем передвижения по синтезированному дереву решений от корневого узла к листу в зависимости от значений входных параметров.

Этап сокращения (усечение, обрезка, pruning) используется для упрощения построенного дерева путем отсечения потомков у выбранной вершины, которая впоследствии становится листом с определенным значением. Усечение узла выполняется в случае, если оно не приведет к существенному ухудшению аппроксимационных и обобщающих характеристик дерева решений.

Таким образом, этап усечения дерева выполняется снизу вверх: движение начинается от листов дерева и происходит вверх до тех пор, пока аппроксимационные способности дерева решений остаются приемлемыми.

### 3. Индукция лингвистических правил на основе построения деревьев решений

Существующие методы построения деревьев решений [2–7] не учитывают особенностей задачи индукции лингвистических правил. В связи с этим разрабатывается новый метод построения деревьев решений для индукции правил. Подобно известным методам построения дерева решений, предлагаемый метод состоит из основных фаз: рост дерева и его сглаживание (сокращение), после чего выполняется преобразование дерева решений в лингвистические правила. Наиболее важными аспектами предлагаемого метода являются следующие: использование модифицированной энтропии как оценочной меры, и использование сглаживания для отсечения.

Таким образом, предлагаемый метод состоит из следующих этапов:

– рост дерева;

– сглаживание дерева;

– преобразование дерева решений в лингвистические правила.

На этапе роста дерева предлагается использовать жадный подход. В каждом узле, соответствующем подмножеству $T$ обучающей выборки, выбирается признак $f$ и значение $v$ таким образом, что данные из $T$ разделяются на два подмножества $T^1_{f,v}$ и $T^1_{f,v}$ исходя из условий $x_{i,f} \le v : T^1_{f,v} = \{x_i \in T : x_{i,f} \le v\}$ и $T^2_{f,v} = \{x_i \in T : x_{i,f} > v\}$. Такое разбиение разделяет множество объектов обучающей выборки на такие, для которых значение признака $f$ меньше значения $v$, и на те, для которых значение признака $f$ больше значения $v$.

С целью разбиения дерева решений для каждого возможного разбиения ($f$, $v$) рассчитывается оценочная функция (1):

$$Q(f,v) = p_{f,v} g(p^1_{f,v}) + (1 - p_{f,v}) g(p^2_{f,v}), \qquad (1)$$

где $p^1_{f,v} = P(y_i = 1 \mid x_i \in T^1_{f,v})$, $p^2_{f,v} = P(y_i = 1 \mid x_i \in T^2_{f,v})$ и $p_{f,v} = P(x_i \in T^1_{f,v} \mid x_i \in T)$; $g(p)$ – модифицированная энтропия для вероятности отнесения выходной переменной $y$ к рассматриваемому классу при условии, что x больше или меньше значения $v$ ( $p^1_{f,v}$ и $p^2_{f,v}$ соответственно):

$$g(p) = -r(p)\ln(r(p)) - (1 - r(p))\ln(1 - r(p)), \qquad (2)$$

где $r(p)$ преобразует оценку вероятности:

$$r(p)=\begin{cases}\frac{1}{2}(1+\sqrt{2p-1}), \text{ если } p>0{,}5;\\ \frac{1}{2}(1-\sqrt{1-2p}), \text{ если } p>0{,}5.\end{cases}, \qquad (3)$$

Таким образом, чем ближе значение вероятности к 0,5, тем выше модифицированное значение, а чем дальше от 0,5, тем значение ниже.

Оценочная функция рассчитывается для всех возможных разбиений и выбирается разбиение с наименьшим значением оценочной функции. Разбиение начинается от корневого узла и продолжается до тех пор, пока не возникнет ситуация, когда невозможно произвести новое разбиение.

После выполнения первого этапа может возникнуть ситуация «переобучения» дерева, что может привести к не совсем корректной работе дерева на тестовых выборках. В связи с этим на втором этапе производится усечение большого дерева, чтобы дерево меньших размеров давало более стабильные оценки вероятности и было более интерпретабельным.

Далее описывается подход, который вместо урезания полного дерева будет производить переоценку вероятности каждого листового узла путем усреднения оценки вероятности по пути следования от корневого узла к листовому узлу. Для достижения данной цели была взята идея «утяжеления дерева» [8]. Если используется дерево для сжатия бинарного классового признака $y_i$, основанного на $x_i$, то в таком случае метод утяжеления дерева гарантирует, что коэффициент сжатия переоцененной вероятности не будет хуже, чем в успешно усечённом дереве. Так как предлагаемый метод применяется в большей степени к трансформированной оценке вероятности $r(p)$, чем непосредственно к $p$, то теоретически результат может быть следующим: путем использования переоцененной вероятности можно достичь ожидаемую классификацию обучающего множества с не худшим результатом, чем у правильно усеченного дерева.

Следует отметить, что данный подход также является сжатием, поскольку при помощи такого подхода оценка сжимается от дальних узлов дерева по направлению к оценкам узлов, которые находятся ближе к корню дерева.

Пусть узлы $T_1$ и $T_2$ являются элементами одного уровня с общим родительским узлом $T$. Пусть $p(T_1)$, $p(T_2)$ и $p(T)$ будут соответствующими оценками вероятности. Локальная переоцененная вероятность это $w_T p(T)+(1-w_T)p(T_1)$ для $T_1$ и $w_T p(T)+(1-w_T)p(T_2)$ для $T_2$. Локальная значимость $w_T$ и сопутствующая функция $G(T)$ рассчитываются рекурсивно, основываясь на следующих формулах:

$$\frac{w_T}{1-w_t}=\frac{c\cdot\exp(-|T|\,g(p(T)))}{\exp(-|T_1|\,G(T_1)-|T_2|\,G(T_2))},$$

$$G(T)=\begin{cases}g(p(T))+\frac{1}{|T|}\log((1+\frac{1}{c})w_T), \text{ если } w_T>0{,}5,\\ \frac{|T_1|}{|T|}G(T_1)+\frac{|T_2|}{|T|}G(T_2)+\\ \quad+\frac{1}{|T|}\log((1+c)(1-iw_T)), \text{ в противном случае.}\end{cases}$$

Параметр $c$ устанавливается априорно и показывает Байесову «оценку» разбиения. Для листового узла $T$ устанавливается: $G(T)=g(p(T))$ и $w_T=1$.

После вычисления значимостей $w_T$ для каждого узла рекурсивным методом (используется нисходящая рекурсия) необходимо рассчитать глобальную оценку вероятности для каждого узла дерева сверху вниз. Данный этап усредняет все оценки $r(p)$ от корневого узла $T_0$ к узлу $T_h$ по пути $T_0$, ..., $T_h$, основываясь на значимости $w_T$. Следует отметить, что значимость $w_T$ является лишь локально важной. Это означает то, что глобальная значимость узла $T_h$ является $w_T{}^*=\prod_{i<k}(1-w_i)w_k$ на всём пути. По определению, $\sum_{i=1}^{h} w_i=1$ для любого направления, ведущему к листу. По следующим рекурсивным формулам вычисляется глобальная переоценка подчиненных узлов $T_1$ и $T_2$ в родительском узле $T$:

$$\hat{w}_{T_i}=\hat{w}_T(1-w_T), \qquad (4)$$

$$r^*(T_i)=r^*(T)+\hat{w}_{T_i}w_{T_i}r(p(T_i)), \qquad (5)$$

где $r(p(T))$ – преобразование по формуле (3) из оценки вероятности $p(T)$ в узле $T$. В корневом узле устанавливается: $\hat{w}=1$. После вычисления $r^*(T_h)$ для листового узла $T_h$ в качестве оценки вероятности можно использовать следующее: $r^{-1}(r(T_h))$. Метка класса для $T_i$ будет равна единице, если $r(T_i)>0{,}5$, в противном случае – нулю. Усечение дерева выполняется, начиная с основания по направлению вверх путем проверки идентичности узлов одного уровня. Если идентичность узлов выявлена, то они удаляются и используется значение родительского узла. Данная процедура будет продолжаться до тех пор, пока она не станет невозможной. Метод сглаживания последовательно улучшает работу дерева. Оценка временной сложности – $O(M)$, где $M$ – количество узлов неусеченного дерева.

Неотъемлемой частью предложенного метода является этап преобразования дерева решений в эквивалентный набор легко поддающихся толкованию лингвистических правил. Важность такого преобразования объясняется двумя причинами.

1. Любому человеку легче понять и изменить набор правил, чем понять и изменить дерево решений. Потребность в таком изменении очевидна. Например, может возникнуть ситуация, когда имеется некоторое несоответствие между обучающей выборкой и реальной системой, что требует

ручной модификации автоматически созданной системы, и, таким образом, в системе, основанной на правилах, такую модификацию можно выполнить путём простого изменения соответствующих правил.

2. Тот факт, что набор правил логически эквивалентен соответствующему дереву решений для данной обучающей выборки, гарантирует, что любой математический анализ эффективности работы дерева решений относится не только к дереву решений, но и к соответствующему набору правил.

Самый простой способ преобразования дерева в эквивалентный набор правил заключается в том, чтобы создать набор правил из правил, каждое из которых соответствует отдельному листу дерева путём формирования логического объединения условий на пути от корня дерева к листу.

Предлагается подход, преобразующий дерево решений в набор логически эквивалентных правил. Целью предлагаемого подхода не является получение доказуемо минимального набора правил. Вместо этого с помощью предлагаемого подхода производится логическое усечение правил.

1. Проверка условий “>” и “<” во всех правилах с целью устранения избыточности в описании условий правил. Таким образом, выполняется, например, следующее преобразование: $(x < 3) \cap (x < 5)$ заменяется на $(x < 3)$.

2. Удаление условий, которые являются логически избыточными в контексте всего набора правил, т.е. удаление условий, которые идентифицируются исходя из структуры полученного дерева решений. Такое упрощение изменяет правило, связанное с конкретным листом дерева, при этом сохраняя полную адекватность всего набора правил.

Для каждого листа, отнесённого к классу *X,* создаётся правило о том, что объект относится к классу *X* путём конъюнкции условий, находящихся на пути следования от корня к *X*, но используя только те условия, которые соответствуют следующему правилу: для каждого узла *N* на пути от корня к листу с меткой *X*, условие, соответствующее родителю *N*, является частью конъюнкции только в том случае, если срабатывает условие соседства для узла *N*. Условие соседства для *N* считается успешным, если: узел *N* не является корнем, и соседний узел относительно *N* не является листом с меткой *X*.

Таким образом, результирующий набор правил является логически эквивалентным базовому дереву решений.

Предложенный метод индукции лингвистических правил с использованием деревьев решений был программно реализован на языке программирования C#.

Для экспериментов использовались тестовые данные, которые были взяты из общедоступных репозиториев [9]. Экспериментальные исследования проводились на основании выборки, которая содержала информацию об эхокардиограммах пациентов с сердечными приступами. Выборка содержала информацию о 132 пациентах, каждый из которых характеризовался 12 признаками. Кроме того, для каждого пациента указывалось, жив он или умер.

Предложенный метод индукции нечётких правил сравнивался с мультиагентным методом и каноническим методом эволюционного поиска. Исходя из проведенных экспериментов, были получены базы лингвистических правил, характеризующиеся следующим качеством классификации пациентов: 81,3%, 79,1% и 92,7% для мультиагентного, эволюционного и предложенного методов, соответственно.

Таким образом, можно отметить, что предложенный метод построения деревьев решений для индукции лингвистических правил обеспечивает более точные результаты прогнозирования по сравнению с другими известными методами индукции лингвистических правил.

**Выводы**

В работе решена актуальная задача индукции лингвистических правил.

Научная новизна работы заключается в том, что разработан новый метод построения деревьев решений, позволяющий выполнять индукцию лингвистических правил, что достигается за счёт введения дополнительных функций преобразования при росте дерева, путём сглаживания дерева решений для его усечения и за счёт введения критерия соседства при преобразовании дерева решений.

Разработанный метод идентификации деревьев решений для индукции лингвистических правил позволяет произвести преобразование и объединение правил, что обеспечивает возможность разработки экспертных систем на основании более логически прозрачных и простых баз лингвистических правил.

Практическая ценность полученных результатов заключается в том, что на основе предложенного метода разработано программное обеспечение, позволяющее производить индукцию баз правил для получения баз лингвистических правил, на основании которых можно создавать экспертные системы с меньшей ошибкой классификации.

**Список литературы: 1.** *Субботін, С. О.* Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень : навч. посібник [Текст] / С. О. Субботін. – Запоріжжя: ЗНТУ, 2008. – 341 с. **2.** *Encyclopedia* of artificial intelligence / Eds.: J. R. Dopico, J. D. de la Calle, A. P. Sierra. – New York : Information Science Reference, 2009. – Vol.

1–3. – 1677 p. **3.** *Rokach L.* Data Mining with Decision Trees. Theory and Applications / L. Rokach, O. Maimon. – London : World Scientific Publishing Co, 2008. – 264 p. **4.** *Quinlan J. R.* Induction of decision trees / J. R. Quinlan // Machine Learning. – 1986. – № 1. – P. 81–106. **5.** *Quinlan J. R.* Decision trees and decision making / J. R. Quinlan // IEEE Transactions on Systems, Man and Cybernetics. – 1990. – № 2 (20). – P. 339–346. **6.** *Liu X.* A decision tree solution considering the decision maker's attitude / X. Liu, Q. Da // Fuzzy Sets and Systems. – 2005. – № 152 (3). – P. 437–454. **7.** *Classification* and regression trees / L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. – California : Wadsworth & Brooks, 1984. – 368 p. **8.** *Willems F. M. J.* The Context Tree Weighting Method: Basic Properties / F. M. J. Willems, Y. M. Shtarkov, T. J. Tjalkens // IEEE Transactions on Information Theory. – 1995. – № 3. – P. 653–664. **9.** UCI Machine Learning Repository [electronic resource] / Center for Machine Learning and Intelligent Systems. – Access mode : http://archive.ics.uci.edu/ml/datasets.html.

***Поступила в редколлегию 20.06.2011***

УДК 004.93

**Індукція лінгвістичних правил з використанням дерев рішень** / Є. О. Гофман, А. О. Олійник, С. О. Субботін // Біоніка інтелекту: наук.-техн. журнал. – 2011. – № 3 (77). – С. 126-130.

Розглядається завдання індукції лінгвістичних правил. Розроблено метод ідентифікації дерев рішень для індукції лінгвістичних правил. Створено програмне забезпечення на основі запропонованого методу.

Бібліогр.: 9 найм.

UDC 004.93

**Linguistic Rules Induction with Decision Trees** / Ye. A. Gofman, A. O. Oliinyk, S. A. Subbotin // Bionics of Intelligense: Sci. Mag. – 2011. – № 3 (77). –P. 126-130.

The problem of linguistic rules induction is considered in this paper. A method of decision trees identification for the linguistic rules induction is developed. The software based on the proposed method is created.

Ref.: 9 items.