УДК 004.93

Гофман Є. О.[1], Олійник А. О.[2], Субботін С. О.[3]

[1]Аспірант Запорізького національного технічного університету
[2]Канд. техн. наук, доцент Запорізького національного технічного університету
[3]Канд. техн. наук, доцент Запорізького національного технічного університету

СКОРОЧЕННЯ БАЗ ЛІНГВІСТИЧНИХ ПРАВИЛ НА ОСНОВІ ДЕРЕВ РОЗВ'ЯЗКІВ

Розглянуто завдання індукції лінгвістичних правил. Розроблено метод ідентифікації дерев розв'язків для індукції лінгвістичних правил. Створено програмне забезпечення на основі запропонованого методу. Проведено експерименти по розв'язанню практичних задач, що дозволило дослідити ефективність запропонованого методу.

Ключові слова: дерево розв'язків, індукція правил, лінгвістичне правило.

ВСТУП

В наш час широке застосування отримали експертні системи, засновані на лінгвістичних правилах [1, 2], які успішно використовуються в різних прикладних областях, зокрема в технічному та медичному діагностуванні, фінансовому менеджменті, розпізнаванні образів, геологічній розвідці, керуванні комп'ютерними мережами, технологічними процесами, аналізі веб-контенту в інтернет та ін. Широке застосування таких систем обумовлене в першу чергу тим, що вони є прозорими й відносно дешевими в реалізації.

Оскільки бази правил в експертних системах часто характеризуються великим обсягом, актуальним є завдання індукції правил, суть якого полягає в тому, що на основі початкового набору правил необхідно сформувати нову базу правил меншого обсягу, яка в достатній мері представляла б початкову базу правил і була б менш надлишковою.

Існують різні методи індукції правил [3], однак ці методи при обробці правил аналізують їхню якість окремо, не розглядаючи та не враховуючи якість усієї бази в цілому, що приводить до одержання неоптимальних баз нечітких правил. Тому актуальною є розробка нових методів індукції правил, які враховували б якість усієї бази знань, а не тільки окремих правил. Для розв'язання даного завдання пропонується створювати дерева розв'язків [3-5], які б після їхньої побудови переводилися в лінгвістичні правила. Вибір дерев розв'язків обгрунтовується їхньою можливістю виявляти неспостережувані зв'язки всередині досліджуваних об'єктів, процесів і систем.

Метою даної роботи є розробка методу індукції лінгвістичних правил з використанням математичного апарата дерев розв'язків.

Для досягнення поставленої мети необхідно розв'язати такі завдання:

– огляд математичного апарату дерев розв'язків;

– приведення основних етапів ідентифікації дерев розв'язків відповідно до розв'язуваного завдання;

– створення правил перетворення дерев розв'язків у лінгвістичні правила;

– порівняння розробленого підходу з існуючими методами індукції лінгвістичних правил.

ПОСТАНОВА ЗАВДАННЯ

Нехай задана база лінгвістичних правил $RB = \{R^1, R^2 ..., R^{RN}\}$, що описує об'єкти навчальної вибірки $O = \{O^1, O^2, ..., O^N\}$. Тоді на основі навчальної вибірки об'єктів O, необхідно сформувати таку базу лінгвістичних правил $RB^* = \{R^1, R^2, ..., R^{RN^*}\}$, $RN^* << RN$, яка забезпечувала б прийнятну якість прогнозування експертної системи, побудованої на основі отриманої бази лінгвістичних правил RB^*:

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

де $Q(RB^*)$ – точність прогнозування або класифікації по базі правил RB^*; $Q_{threshold}$ – мінімально припустима точність прогнозування або класифікації.

ДЕРЕВА РОЗВ'ЯЗКІВ

Дерева розв'язків являють собою графові інтелектуальні моделі, у внутрішніх вузлах яких розташовані функції прийняття рішень на основі значень вхідних змінних, а в зовнішніх вузлах (термінальних вузлах, листах) знаходяться значення вихідної змінної, відповідні до умов у внутрішніх вузлах [2, 6, 7].

Завдяки своїй деревоподібній структурі такі моделі дозволяють наочно представляти результати обчислень. Тому вони добре інтерпретуються людьми-фахівцями в прикладних областях, які, як правило, не мають спеціальної математичної підготовки та не знайомі з методами й моделями штучного інтелекту. Дерева розв'язків дозволяють ефективно вирішувати завдання класифікації та прогнозування, забезпечуючи при цьому високу точність.

Для застосування дерев розв'язків на практиці з метою класифікації або прогнозування значень вихідних параметрів досліджуваних об'єктів по наборах значень вхідних характеристик необхідно за допомогою даних навчальної вибірки сформувати дерево розв'язків таким чином, щоб воно щонайкраще описувало досліджуваний об'єкт.

Побудова дерев розв'язків пов'язана з витягом правил з навчальних вибірок. Кожний шлях від кореня дерева до одного з його листів може бути перетворений до логічного висловлення – правилу типу «якщо А, то В», де його антецедент виходить шляхом використання всіх умов, представлених у внутрішніх вузлах від кореня до вихідного листа, а права частина правила виходить із відповідного листа дерева.

Процес побудови дерева розв'язків, як правило, містить такі етапи: розростання, розгалуження, обчислення значення вихідного параметра для листа, скорочення.

У результаті етапу розростання (збільшення, growing) деяка вершина заміняється піддеревом, отриманим шляхом розгалуження цієї вершини. На даному етапі відбувається поділ обраної вершини на деякі нові (у випадку дихотомічного дерева вершина розбивається на дві нові). При цьому перебираються всі ознаки й усі можливі варіанти розгалуження по кожній з ознак. У результаті залишається варіант розбиття, при якому значення критерію якості розбиття є найкращим. Якщо нові вершини є перспективними для наступного поділу (критерії завершення розростання не задоволені), то виконується їхнє розгалуження. У випадку неможливості подальшого поділу вершини вона стає листом, і для неї виконується процедура обчислення значення вихідного параметра. Якщо розгалуження вершини приводить до погіршення якості дерева, то вершина також оголошується листом.

Процедура розгалуження (поділу, splitting) дерева викликається рекурсивно при виконанні етапу розростання. Розгалуження призначено для створення для обраної вершини заданої кількості (для дихотомічних дерев – дві) вершин-нащадків.

Обчислення значення вихідного параметра відбувається шляхом пересування по синтезованому дереву розв'язків від кореневого вузла до листа в залежності від значень вхідних параметрів.

Етап скорочення (усікання, pruning) використовується для спрощення побудованого дерева шляхом відсікання нащадків у обраній вершини, яка в наслідку стає листом з певним значенням. Усікання вузла виконується у випадку, якщо воно не приведе до істотного погіршення апроксимаційних і узагальнюючих характеристик дерева розв'язків.

Таким чином, етап усікання дерева виконується знизу нагору: рух починається від листів дерева та відбувається нагору доти, доки апроксимаційні здатності дерева розв'язків залишаються прийнятними.

ІНДУКЦІЯ ЛІНГВІСТИЧНИХ ПРАВИЛ НА ОСНОВІ ПОБУДОВИ ДЕРЕВ РОЗВ'ЯЗКІВ

Існуючі методи побудови дерев розв'язків [2–7] не враховують особливостей завдання індукції лінгвістичних правил. У зв'язку з цим розробляється новий метод побудови дерев розв'язків для індукції правил. Подібно відомим методам побудови дерев розв'язків, пропонований метод складається з основних фаз: ріст дерева і його згладжування (скорочення), після чого виконується перетворення дерева розв'язків у лінгвістичні правила. Найбільш важливими аспектами пропонованого методу є наступні: використання модифікованої ентропії як оцінної міри і використання згладжування для відсікання.

Таким чином, пропонований метод складається з таких етапів:

– ріст дерева;

– згладжування дерева;

– перетворення дерева розв'язків у лінгвістичні правила.

На етапі росту дерева пропонується використовувати жадібний підхід. У кожному вузлі, що відповідає підмножині T навчальної вибірки, вибирається ознака f і значення v таким чином, що дані з T розділяються на дві підмножини $T^1_{f,v}$ та $T^2_{f,v}$ виходячи з умов $x_{i,f} \le v : T^1_{f,v} = \{x_i \in T : x_{i,f} \le v\}$ і $T^2_{f,v} = \{x_i \in T : x_{i,f} > v\}$. Таке розбиття розділяє множину об'єктів навчальної вибірки на такі, для яких значення ознаки f менше значення v, і на ті, для яких значення ознаки f більше значення v.

З метою розбиття дерева розв'язків для кожного можливого розбиття (f, v) розраховується оціночна функція:

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

де $p^1_{f,v} = P(y_i = 1 \mid x_i \in T^1_{f,v})$, $p^2_{f,v} = P(y_i = 1 \mid x_i \in T^2_{f,v})$ і $p_{f,v} = P(x_i \in T^1_{f,v} \mid x_i \in T)$; $g(p)$ – модифікована ентропія для ймовірності віднесення вихідної змінної y до розглянутого класу за умови, що x більше або менше значення v ($p^1_{f,v}$ і $p^2_{f,v}$, відповідно):

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

де $r(p)$ перетворить оцінку ймовірності:

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

Таким чином, чим ближче значення ймовірності до 0,5, тим вище модифіковане значення, а чим далі від 0,5, тим значення нижче.

Функція оцінки розраховується для всіх можливих розбиттів і вибирається розбиття з найменшим значенням оціночної функції. Розбиття починається від кореневого вузла та триває доти, поки не виникне ситуація, коли неможливо зробити нове розбиття.

Після виконання першого етапу може виникнути ситуація «перенавчання» дерева, що може привести до не зовсім коректної роботи дерева на тестових вибірках. У зв'язку із цим на другому етапі проводиться усікання великого дерева, щоб дерево менших розмірів давало більш стабільні оцінки ймовірності й було більш інтерпретабельним.

Далі описується підхід, який замість урізання повного дерева, буде робити переоцінку ймовірності кожного листового вузла шляхом усереднення оцінки ймовірності по шляху проходження від кореневого вузла до листового. Для досягнення даної мети була взята ідея «обважнення дерева» [8]. Якщо використовується дерево для стиснення бінарної класової ознаки y_i, заснованого на x_i, то в такому випадку метод обважнення дерева гарантує, що коефіцієнт стиснення переоціненої ймовірності не буде гірше, чим в успішно усіченому дереві. Оскільки пропонований метод застосовується більшою мірою до трансформованої оцінки ймовірності $r(p)$, ніж безпосередньо до p, то теоретично, результат може бути наступним: шляхом використання переоціненої ймовірності, можна досягнути очікуваної класифікації навчальної множини з не гіршим результатом, ніж у правильно усіченого дерева.

Слід зазначити, що даний підхід також є стисненням, оскільки за допомогою такого підходу оцінка стискується від далеких вузлів дерева в напрямку до оцінок вузлів, які перебувають ближче до кореня дерева.

Нехай вузли T_1 та T_2 є елементами одного рівня із загальним батьківським вузлом T. Нехай $p(T_1)$, $p(T_2)$ і $p(T)$ будуть відповідними оцінками ймовірності. Локальна переоцінена ймовірність може бути обчислена за формулами: $w_T p(T) + (1 - w_T) p(T_1)$ для T_1 та $w_T p(T) + (1 - w_T) p(T_2)$ для T_2. Локальна значимість w_T і супутня функція $G(T)$ розраховуються рекурсивно, грунтуючись на таких формулах:

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

$$G(T) = \begin{cases} g(p(T)) + \dfrac{1}{|T|}\log((1 + \dfrac{1}{c})w_T), \text{ якщо } w_T > 0{,}5, \\ \dfrac{|T_1|}{|T|} G(T_1) + \dfrac{|T_2|}{|T|} G(T_2) + \dfrac{1}{|T|}\log((1 + c)(1 - iw_T)), \text{ в іншому випадку.} \end{cases}$$

Параметр c установлюється апріорно та показує Байєсову «оцінку» розбиття. Для листового вузла T установлюється: $G(T) = g(p(T))$ та $w_T = 1$.

Після обчислення значимостей w_T для кожного вузла рекурсивним методом (використовується спадна рекурсія), необхідно розрахувати глобальну оцінку ймовірності для кожного вузла дерева зверху вниз. Даний етап усереднює усі оцінки $r(p)$ від кореневого вузла T_0 до вузла T_h по шляху $T_0, \ldots, T_h$, грунтуючись на значимості w_T. Слід зазначити, що значимість w_T є лише локально важливою. Це означає те, що глобальна значимість вузла T_h є $w_T^* = \prod_{i<k}(1 - w_i) w_k$ на всьому шляху. За визначенням, $\sum_{i=1}^{h} w_i = 1$ для будь-якого напрямку, що веде до листа. За наступними рекурсивними формулами обчислюється глобальна переоцінка підлеглих вузлів T_1 і T_2 в батьківському вузлі T:

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

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

де $r(p(T))$ – перетворення з оцінки ймовірності $p(T)$ у вузлі T. У кореневому вузлі встановлюється: $\hat{w} = 1$. Після обчислення $r^*(T_h)$ для листового вузла T_h в якості оцінки ймовірності можна використовувати $r^{-1}(r(T_h))$. Мітка класу для T_i буде дорівнювати одиниці, якщо $r(T_i) > 0{,}5$, в іншому випадку – нулю. Усікання дерева виконується, починаючи з основи за напрямом вгору шляхом перевірки ідентичності вузлів одного рівня. Якщо ідентичність вузлів виявлена, то вони видаляються й використовується значення батьківського вузла. Дана процедура буде тривати доти, поки вона не стане неможливою. Метод згладжування послідовно поліпшує роботу дерева. Оцінка часової складності – $O(M)$, де M – кількість вузлів неусіченого дерева.

Важливою частиною запропонованого методу є етап перетворення дерева розв'язків в еквівалентний набір лінгвістичних правил, що легко піддаються тлумаченню. Важливість такого перетворення пояснюється двома причинами:

1. Будь-якій людині легше зрозуміти й змінити набір правил, ніж зрозуміти й змінити дерево розв'язків. Потреба в такій зміні очевидна. Наприклад, може виникнути ситуація, коли є деяка невідповідність між навчальною вибіркою й реальною системою, що вимагає ручної модифікації автоматично створеної системи, і, таким

чином, у системі, заснованій на правилах, таку модифікацію можна виконати шляхом простої зміни відповідних правил.

2. Той факт, що набір правил є логічно еквівалентним відповідному дереву розв'язків для даної навчальної вибірки, гарантує, що будь-який математичний аналіз ефективності роботи дерева розв'язків відноситься не тільки до дерева розв'язків, але й до відповідного набору правил.

Найпростіший спосіб перетворення дерева в еквівалентний набір правил полягає в тому, щоб створити набір правил із правил, кожне з яких відповідає окремому листу дерева шляхом формування логічного об'єднання умов на шляху від кореня дерева до листа.

Пропонується підхід, що перетворить дерево розв'язків у набір логічно еквівалентних правил. Метою пропонованого підходу не є одержання доказово мінімального набору правил. Замість цього за допомогою пропонованого підходу проводиться логічне усікання правил.

1. Перевірка умов «>» та «<» у всіх правилах з метою усунення надмірності в описі умов правил. Таким чином, виконується, наприклад таке перетворення: $(x < 3) \cap (x < 5)$ заміняється на $(x < 3)$.

2. Видалення умов, які є логічно надлишковими в контексті всього набору правил, тобто видалення умов, які ідентифікуються виходячи зі структури отриманого дерева розв'язків. Таке спрощення змінює правило, що пов'язано з конкретним листом дерева, при цьому зберігаючи повну адекватність усього набору правил.

Для кожного листа, що віднесений до класу X, створюється правило про те, що об'єкт відноситься до класу X шляхом кон'юнкції умов, що знаходяться на шляху проходження від кореня до X, але використовуючи тільки ті умови, які відповідають наступному правилу: для кожного вузла N на шляху від кореня до листа з міткою X умова, що відповідає батьку N, є частиною кон'юнкції тільки в тому випадку, якщо спрацьовує умова сусідства для вузла N. Умова сусідства для N вважається успішною, якщо: вузол N не є коренем і сусідній вузол відносно N не є листом з міткою X.

Таким чином результуючий набір правил є логічно еквівалентним базовому дереву розв'язків.

Запропонований метод індукції лінгвістичних правил з використанням дерев розв'язків був програмно реалізований за допомогою мови програмування C#.

Для експериментів використовувалися тестові дані, які були взяті із загальнодоступних репозиторіїв [9]. Експериментальні дослідження проводилися на підставі вибірки, яка містила інформацію про ехокардіограми пацієнтів із серцевими приступами. Вибірка містила інформацію про 132 пацієнтів, кожний з яких характеризувався 12 ознаками. Крім того, для кожного пацієнта вказувалося живий він або помер.

Запропонований метод індукції нечітких правил порівнювався з мультиагентним методом і канонічним методом еволюційного пошуку. Виходячи із проведених експериментів, були отримані бази лінгвістичних правил, що характеризуються наступною якістю класифікації пацієнтів: 81,3 %, 79,1 % і 92,7 % для мультиагентного, еволюційного та запропонованого методів, відповідно.

Таким чином, можна відзначити, що запропонований метод побудови дерев розв'язків для індукції лінгвістичних правил забезпечує більш точні результати прогнозування в порівнянні з іншими відомими методами індукції лінгвістичних правил.

ВИСНОВКИ

У роботі вирішено актуальне завдання автоматизації індукції лінгвістичних правил.

Наукова новизна роботи полягає в тому, що розроблено новий метод побудови дерев розв'язків, який дозволяє виконувати індукцію лінгвістичних правил, що досягається за рахунок введення додаткових функцій перетворення при рості дерева, шляхом згладжування дерева розв'язків для його усікання та за рахунок введення критерію сусідства при перетворенні дерева розв'язків.

Розроблений метод ідентифікації дерев розв'язків для індукції лінгвістичних правил дозволяє виконувати перетворення й об'єднання правил, що забезпечує можливість розробки експертних систем на підставі більш логічно прозорих і простих баз лінгвістичних правил.

Практична цінність отриманих результатів полягає в тому, що на основі запропонованого методу розроблено програмне забезпечення, яке дозволяє виконувати індукцію баз правил для одержання баз лінгвістичних правил, на підставі яких можна створювати експертні системи з меншою помилкою класифікації.

СПИСОК ЛІТЕРАТУРИ

1. 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.
2. *Барсегян, А. А.* Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP : учебное пособие / А. А. Барсегян. – С. Пб. : BHV, 2007. – 384 с.
3. *Quinlan, J. R.* Decision trees and decision making / J. R. Quinlan // IEEE Transactions on Systems, Man and Cybernetics. – 1990. – № 2 (20). – P. 339–346.
4. Quinlan J. R. Induction of decision trees / J. R. Quinlan // Machine Learning. – 1986. – № 1. – P. 81–106.
5. Gelfand S. B. An Iterative Growing and Pruning Algorithm for Classification Tree Design / S. B. Gelfand, C. S. Ravishankar, E. J. Delp // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1991. – № 13 (2). – P. 163–174.
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.

Стаття надійшла до редакції 28.12.2011.

Гофман Е. А., Олейник А. А., Субботин С. А.

СОКРАЩЕНИЕ БАЗ ЛИНГВИСТИЧЕСКИХ ПРАВИЛ НА ОСНОВЕ ДЕРЕВЬЕВ РЕШЕНИЙ

Рассмотрена задача индукции лингвистических правил. Разработан метод идентификации деревьев решений для индукции лингвистических правил. Создано программное обеспечение на основе предложенного метода. Проведены эксперименты по решению практических задач, что позволило исследовать эффективность предложенного метода.

Ключевые слова: дерево решений, индукция правил, лингвистическое правило.

Gofman Ye., Oliinyk A., Subbotin S.

LINGUISTIC RULES BASES REDUCTION BASED ON DECISION TREES

The problem of linguistic rules induction is considered. A method of decision trees identification for linguistic rules induction is developed. The software based on the proposed method is created. Experiments on the solution of practical problems, which allowed to investigate the effectiveness of the proposed method are made.

Key words: decision tree, rules induction, linguistic rule.