4. Каплин Е. А. Принципы построения и основы функционирования пакетных радиосетей в нестационарных средах передачи сообщений. Электросвязь. 1994. № 10. С. 62–70.

DOI https://doi.org/10.30525/978-9934-588-79-2-1.18

ОСОБЛИВОСТІ РЕАЛІЗАЦІЙ МОДЕЛЕЙ ДЕРЕВ КЛАСИФІКАЦІЙ НА ОСНОВІ СЕЛЕКЦІЇ ОЗНАК

Повхан І. Ф.
кандидат технічних наук, доцент,
доцент кафедри програмного забезпечення систем
Ужгородського національного університету
м. Ужгород, Україна

Вступ Відмітимо, що станом на сьогоднішній день відомо біля трьох десятків готових програмних систем (ПС) для побудови різних типів моделей дерев класифікації (дерев рішень) у вигляді структур ЛДК (RStudio, RulQuest, DTTL v1.5, RLQTree, DCT v7, Precision Tree System, Edraw, SHAIDS, Weka, ЛАСТАН, АУРОН та інші) та лише одна ПС яка базується на концепції АДК (ОРІОН). Всі ці системи відрізняються прикладною спрямованістю задач що розв'язуються, методами та концептуальними засадами, різноманітним рівнем підтримки, причому багато з них знаходять у вільному (або частково вільному) доступі. Домінуючими підходами є системи на основі методів CART (спрямованих для розв'язку задач класифікації та регресивного аналізу), а також ПС на основі схеми C4.5/C5.0 та її сучасних модифікації (для розв'язку задач розпізнавання та класифікації) та ID3.

Основна частина. Відмітимо, що алгоритм ID3 є однією з найпростіших схем для отримання дерев рішень з категоріальними класами та атрибутами (на основі ентропійного критерію), причому саме на основі ID3 і був написаний пізніше алгоритм C4.5. Хоча існує достатньо багато реалізаційних схем різних методів та підходів дерев рішень (дерев класифікації), одною з найбільш вживаних та такою що забезпечує необхідну ефективність є алгоритмічна схема C5.0 (подальший розвиток концепції алгоритму C4.5), причому вона стала по факту галузевим стандартом для побудови моделей дерев класифікації – оскільки підходить для більшості типів задачах безпосередньо прикладного характеру в сегменті аналізу різнотипної інформації. В порівнянні з більш досконалими та складними моделями машинного на-

вчання (наприклад – концепцією нейронних мереж) дерева класифікації в рамках схеми C5.0 зазвичай забезпечують не гіршу ефективність та точність, але в той самий час більш підходять для зовнішнього аналізу та фінальної покрокової інтерпретації (виділенні правил класифікації), тобто є достатньо простими та зрозумілими для експерта та спостерігача.

Зауважимо що ПС, які базуються на алгоритмах схеми C5.0 (за авторством J. Ross Quinlan) використовують в якості критерію чистоти підмножин початкової навчальної вибірки (НВ) параметр ентропії. Використовуючи ентропію в якості міри чистоти (однорідності) класів (підмножин початкової НВ), які є результатом процедури розбиття (розгалуження структури дерева), алгоритм може зафіксувати (відібрати) ту ознаку (атрибут), розбиття за якою дає саму чисту (однорідну) підмножини початково НВ (тобто підмножину початкової НВ з найменшою ентропією). Дана схема в літературі позначається – *information gain* (схема підсилення інформації), причому якщо для відібраної ознаки X_i величина *information gain* є нульовою, то це фактично означає безперспективність (неможливість) розбиття НВ на підмножини – не приводить до зменшення коефіцієнту ентропії. Підкреслимо, що максимально можливе значення величини *information gain* дорівнює величині ентропії до розбиття, а це в свою чергу означає, що ентропія після поточного розбиття частини НВ буде дорівнювати нулю для повністю чистих (однорідних) підмножин початкової НВ.

В зв'язку з тим, що структури логічних дерев класифікації (ЛДК) після побудови за вибірками реальних даних великого об'єму – мають в більшості випадків складну для аналізу та неоднорідну за рівнями (ярусами) структуру, то принциповою проблемою залишається питання організації процедури оптимізації або обрізки (*pruning*) таких конструкції. Під складністю структури ЛДК розуміється загальні кількість вершин конструкції дерева (вузлів розгалуження), в такому випадку зазвичай мається на увазі, що модель ЛДК перевизначена (*is overfitted*). Важливою особливістю схеми C5.0 в плані корекції структури побудованого дерева є можливість використання механізму – *post-pruning*, коли відкидаються ті вузли, блоки конструкції, піддерева які мало (відповідно деякого заданого критерію) впливають на результат загальної класифікації (допустима помилка), причому допускається не лише просте відсікання структур дерева, але і їх перенесення в іншу частину структури ЛДК, або заміну на іншу конструкцію з меншою структурною складністю (меншої розгалуженості). Дані схеми оптимізації (обрізки) структур ЛДК – *subtree raising* (підняття піддерева) та *subtree replacement* (заміна піддерева) в процедурі *prun-*

ing використовуються в C5.0 та в небагатьох інших методах побудови дерев класифікації, причому абсолютна більшість інших методів та схем базується на процедурі попередньої обрізки структури ЛДК що будуються – *pre-pruning*, яка має суттєві недоліки в плані можливості пропуску (відсікання) важливих даних, які важко виявити. Так в Ужгородському національному університеті на основі методу розгалуженого вибору ознак ЛДК (селекції наборів елементарних ознак) була розроблена ПС DeTree яка базується на концепції розгалуженого вибору ознак та дозволяє працювати з НВ великого та надвеликого об'єму. Так для перевірки побудованого програмного забезпечення використовувалась відома задача про тип лісового покриву (ліс – 581012 елементів масиву вибірки). Структура вибірки (НВ та ТВ) містить сім класів розбиття (типи можливого лісового покрову), причому об'єкт класифікації представляється як послідовність 12 числових ознак та додатково двох багатозначних дискретних атрибутів. Зауважимо що половина початкового масиву вибірки (а саме 290506 об'єктів відомої класифікації) відводилась для навчання системи, а інша частина для тестування побудованих моделей ЛДК. Загальні дані про задачу та безпосередньо сам масив вибірки для перевірки можна отримати з ресурсу *UCI KDD Archive* (http://kdd.ics.uci.edu).

Основні результати тестування приведені в наступних порівняльних таблицях – (Табл. 1 – Табл. 3).

Таблиця 1

Порівняння схем побудови ЛДК за кількістю помилок класифікації, кількістю правил класифікації, та часом побудови ЛДК

Алгоритмічна схема	*Загальна кількість помилок* - Er_{Alt}	*Загальна кількість правил класифікації* - Rt_{Alt}	*Загальний час генерації ЛДК* - T_{Alt}
C4.5	7.2%	5420	Config. №1 – 34 с. Config. №2 – 42 с.
C5.0	6.3%	4845	Config. №1 – 186 с. Config. №2 – 230 с.
DeTree	6.7%	5028	Config. №1 – 102 с. Config. №2 – 129 с.

Таблиця 2

Порівняння схеми бустингу для алгоритму C5.0

Первинне ЛДК алгоритму C5.0	*ЛДК алгоритму C5.0 на основі бустингу*	*Набір первинних класифікаторів алгоритму C5.0*	*Набір класифікаторів алгоритму C5.0 на онові бустингу*
6.7%	3.8%	6.2%	3.6%

Таблиця 3

Порівняння схем побудови ЛДК за фіксованою точністю, кількістю вузлів та часом побудови структури ЛДК

Алгоритмічна схема	*Загальна кількість помилок - Er_{All}*	*Загальна кількість вузлів ЛДК - Vt_{All}*	*Загальний час генерації ЛДК - T_{All}*
C4.5	6.8%	10167	Config. №1 – 57 с. Config. №2 – 43 с.
C5.0	6.8%	9201	Config. №1 – 62 с. Config. №2 – 51 с.
DeTree	6.7%	1012	Config. №1 – 50 с. Config. №2 – 47 с.

Висновки. Отже зважаючи на все вище сказане, можна зафіксувати наступні принципові моменти:

1) В зв'язку з тим, що структури дерев класифікації після побудови за вибірками реальних даних великого об'єму – мають в більшості випадків складну для аналізу та неоднорідну за рівнями (ярусами) структуру, то принциповою проблемою залишається питання організації процедури оптимізації або обрізки (*pruning*) таких конструкції.

2) Відмітимо, що не дивлячись на високу ефективність в практичній площині, наявність якісного механізму оптимізації, мінімізації побудованих структур дерев класифікації схема C5.0 не позбавлена і певних системних недоліків, які обов'язково потрібно враховувати як при реалізації так і при роботі з побудованими моделями ЛДК.

3) ПС DeTree базується на концепції розгалуженого вибору ознак (поетапної селекції ознак) та дозволяє працювати з НВ різнотипної інформації широкого спектру прикладних задач.

4) Послідовність кроків побудови структури дерева класифікації (моделі ЛДК) в ПС DeTree можна визначити наступним порядком дій:

початковий етап визначення та ініціалізації базових параметрів, етап валідації вхідних даних та визначення режимів роботи ПС, етап роботи базових процедур менеджера пам'яті, етап формування та інформаційної оцінки наборів ознак (атрибутів), етап формування структури ЛДК (вузлів та переходів), етап оптимізації та мінімізації конструкції ЛДК (фінальної обрізки дерева класифікації), етап кінцевої перевірки параметрів побудованої моделі класифікації (ЛДК), етап аналізу та виділення правил класифікації.

Література:

1. Повхан І.Ф., Лавер В.О. Алгоритми побудови логічних дерев класифікації в задачах розпізнавання образів // *Вчені записки Таврійського національного університету*. 2019. Серія: технічні науки. Том 30 (72). № 4 2019. С. 192–201.

2. Повхан І.Ф., Василенко Ю.А., Василенко Е.Ю. Концептуальна основа систем розпізнавання образів на основі метода розгалуженого вибору ознак // *Науково технічний журнал «European Journal of Enterprise Technologies».* 2004. № 7[1]. С. 13–15.

3. Повхан І.Ф. Проблема функціональної оцінки навчальної вибірки в задачах розпізнавання дискретних об'єктів. // *Вчені записки Таврійського національного університету.* 2018. Серия: технічні науки. Том 29 (68) № 6 2018. С. 217–222.

4. Povhan I. Designing of recognition system of discrete objects. // *IEEE First International Conference on Data Stream Mining & Processing (DSMP).* Lviv – 2016, Ukraine, P. 226–231.

5. Повхан І.Ф. Метод розгалуженого вибору ознак в математичному конструюванні багаторівневих систем розпізнавання образів // *Науково технічний журнал «Штучний Інтелект».* 2003. № 7. С. 246–249.

6. Povhan I. General scheme for constructing the most complex logical tree of classification in pattern recognition discrete objects. // *Збірник наукових праць "Електроніка та інформаційні технології",* Львів. – 2019. – Випуск 11. – С. 112–117.

7. Повхан І.Ф. Задача апроксимації вибірки дискретних наборів геометричними об'єктами. // *Вчені записки Таврійського національного університету*. Серия: технічні науки. – 2019. – Том 30 (69) № 3. 2019. – С. 136–142.