Keywords: cryptography, asymmetric cryptosystem, vulnerability, secret key, public key, Euler function, lemma, prime, even number, odd number, Fermat numbers.

DOI: 10.18469/ikt.2015.13.4.17

Alekseev Aleksander Petrovich, PhD in Technical Science, Professor of the Department of Information and Computer Engineering, Povolzhskiy State University of Telecommunications and Informatics, Samara, Russian Federation. Tel. +78462280057. E-mail: apa2008@rambler.ru

References

1. Rivest R., Shamir A., Adleman L. A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. *Communications of the ACM*, 1978, vol. 21, no. 2, pp. 120-126. doi: 10.1145/359340.359342
2. Diffie, W., Hellman, M. New Directions in Cryptography. *IEEE Trans. Inform. Theory IT-22*, 1976, pp. 644-654. doi: 10.1109/TIT.1976.1055638
3. Alekseev A.P., Orlov V.V. *Steganograficheskie i kriptograficheskie metody zashchity informatsii: uchebnoe posobie* [Steganographic and cryptographic methods of information protection]. Samara, PGUTI Publ., 2010. 330 p.
4. Alekseev A.P. *Informatika dlya kriptoanalitikov: uchebnoe posobie* [Informatics for cryptanalysts]. Samara, PGUTI Publ., 2015. 376 p.
5. Alekseev A.P. *Informatika 2015* [Informatics 2015]. Moscow, SOLON-Press Publ., 2015. 400 p.
6. RSA. Available at: https://ru.wikipedia.org/wiki/RSA. (Accessed 1.08.2015).
7. Shnajer B. *Prikladnaja kriptografija. Protokoly, algoritmy, ishodnye teksty na jazyke Si* [Applied Cryptography. Protocols, algorithms, source code in C]. Moscow, TRIUMF Publ., 2002. 816 p.
8. Smart N. *Kriptografija* [Cryptography]. Moscow, Tehnosfera Publ., 2006. 528 p.
9. Jashhenko V.V. *Vvedenie v kriptografiju* [Introduction to Cryptography]. St. Petersburg, Piter Publ., 2001. 288 p.
10. Fergjuson N., Shnajer B. *Prakticheskaja kriptografija* [Practical Cryptography]. Moscow, Izdatelskij dom «Viljam», 2005. 424 p.
11. Alferov A.P., Zubov A.Ju., Kuzmin A.S., Cheremushkin A.V. *Osnovy kriptografii* [Basics of cryptography]. Moscow, Gelios ARV Publ., 2002. 480 p.
12. Rjabko B.Ja., Fionov A.N. *Osnovy sovremennoj kriptografii i steganografii* [The foundations of modern cryptography and steganography]. Moscow, Gorjachaja linija-Telekom Publ., 2010. 232 p.
13. Romanec Ju.V., Timofeev P.A., Shangin V.F. *Zashhita informacii v kompjuternyh sistemah i setjah* [Protecting information in computer systems and networks]. Moscow, Radio i svjaz Publ., 2001. 376 p.

Received 10.08.2015

УПРАВЛЕНИЕ И ПОДГОТОВКА КАДРОВ ДЛЯ ОТРАСЛИ ИНФОКОММУНИКАЦИЙ

УДК 004.8

СРАВНЕНИЕ КЛАССИФИКАЦИОННЫХ ВОЗМОЖНОСТЕЙ АЛГОРИТМОВ C4.5 И C5.0

Пальмов С.В., Мифтахова А.А.
Поволжский государственный университет телекоммуникаций и информатики, Самара, РФ
E-mail: psv@psuti.ru

В статье проводится сравнение возможностей алгоритмов деревьев решений C4.5 и C5.0 - одних из наиболее эффективных инструментов классификации интеллектуального анализа данных. Для этого были выбраны две их программные реализации – отечественная аналитическая платформа Deductor и система See5. Чтобы повысить качество сравнительного анализа, использовались три разных набора данных. Как показали результаты эксперимента, утверждения автора-разработчика обоих алгоритмов Куинлана о том, что новая версия алгоритма во всем превосходит старую, оказались несколько излишне оптимистичными. C5.0 действительно строит, как и заявлено, более компактные деревья решений, но скорость его работы осталась сопоставимой с C4.5, а достоверность

получаемой классификационной модели снизилась. Однако авторы статьи не исключают, что вышеуказанные результаты объясняются, тем, что в их распоряжении имелась демонстрационная версия системы See5, которая может обрабатывать файлы, содержащие не более 400 записей.

Ключевые слова: Data Mining, дерево решений, C4.5, C5.0, Deductor, See5.

Введение

Существуют различные подходы к анализу данных [1]. Одним из самых известных является технология интеллектуального анализа данных (Data Mining, далее DM) – процесс обнаружения в необработанных данных ранее неизвестных нетривиальных знаний, необходимых для принятия решений в различных сферах человеческой деятельности [2]. Технология DM располагает большим числом инструментов (алгоритмов) для проведения различных видов анализа [3]. Одним из наиболее популярных классификационных алгоритмов являются деревья решений. Деревья решений – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение [4].

Виды алгоритмов деревьев решений

Наиболее эффективным алгоритмом деревьев решений считается C4.5 – усовершенствованная версия алгоритма ID3 [5], разработанная Д. Куинланом, позволяющая строить дерево решений с неограниченным числом ветвей у узла [6]. Однако некоторое время назад появилась новая модификация – C5.0. Как утверждает автор (все тот же Д. Куинлан), она превосходит предыдущую версию: работает быстрее, строит деревья решений меньшей размерности, использование памяти компьютера является более эффективным, имеет более высокую точность результатов, позволяет автоматически удалять незначащие атрибуты [7]. Для проверки данного утверждения нами были выбраны два программных продукта, в которых реализованы алгоритмы C4.5 и C5.0.

Программные продукты

Алгоритм C4.5. реализован в системе Deductor. Аналитическая платформа Deductor – основа для создания законченных прикладных решений. Реализованные в Deductor технологии позволяют на базе единой архитектуры пройти все этапы построения аналитической системы: от создания хранилища данных до автоматического подбора моделей и визуализации полученных результатов [8-9].

Алгоритм C5.0 реализован в системе See5 [10]. See5 – инструмент анализа данных для прогнозирования диагностического класса какого-либо объекта по значениям его признаков. Содержит единственный обработчик – дерево решений. Причина выбора именно этих программных систем заключается в наличии бесплатных демонстрационных версий, доступных для свободного скачивания с сайтов компаний-разработчиков.

Данные

Для проведения сравнительного анализа было выбрано три набора данных: файл «ирисы Фишера» [11]; файл, содержащий информацию о результатах голосования депутатов Конгресса США (входит в дистрибутив аналитической платформы Deductor); файл, содержащий реальную информацию об абитуриентах одного из вузов РФ. Структура файлов:

- «Ирисы Фишера» – 150 записей, 5 атрибутов (целевой атрибут – «класс»);
- голосование депутатов – 400 записей, 17 атрибутов (целевой атрибут – «партийная принадлежность»);
- абитуриенты – 400 записей, 20 атрибутов (целевой атрибут – «форма обучения»).

Описание эксперимента

Эксперимент состоял из трех частей: построение дерева решений для файла «Ирисы Фишера» средствами Deductor и See5, построение дерева решений для файла «Голосование депутатов» средствами Deductor и See5, построение дерева решений для файла «Абитуриенты» средствами Deductor и See5. Для всех трех случаев были выбраны следующие настройки обработчиков: уровень доверия, используемый при отсечении узлов дерева, – 20%, минимальное количество примеров в узле, при котором будет создан новый, – 2. Результаты эксперимента приведены на рис. 1-6 и в таблице 1.

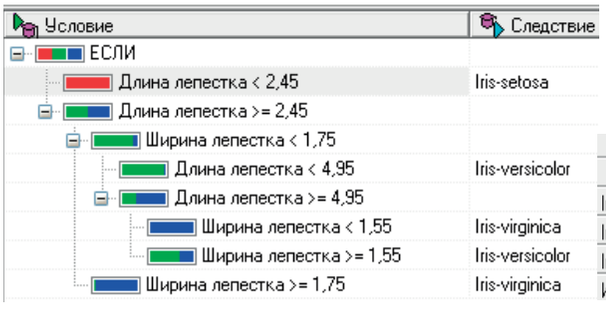
Выводы

Как видно из представленных результатов, во всех трех случаях алгоритм C5.0 построил более компактные деревья, содержащие, как следствие, и меньшее количество правил (см. таблицу 1). Однако точность классификации у алгоритма C4.5 (см. таблицы сопряженности) оказалась несколько выше. Скорость генерации результатов

```
Decision tree:

Dlina lepestka <= 1.9: Iris-setosa (50)
Dlina lepestka > 1.9:                                  (a)   (b)   (c)    <-classified as
:...Shirina lepestka > 1.7: Iris-virginica (46/1)     ----  ----  ----
    Shirina lepestka <= 1.7:                           50                  (a): class Iris-setosa
    :...Dlina lepestka <= 4.9: Iris-versicolor (48/1)        47     3     (b): class Iris-versicolor
        Dlina lepestka > 4.9: Iris-virginica (6/2)            1    49     (c): class Iris-virginica
```

Рис. 1. Дерево решений и таблица сопряжённости для «Ирисов Фишера» (See5)



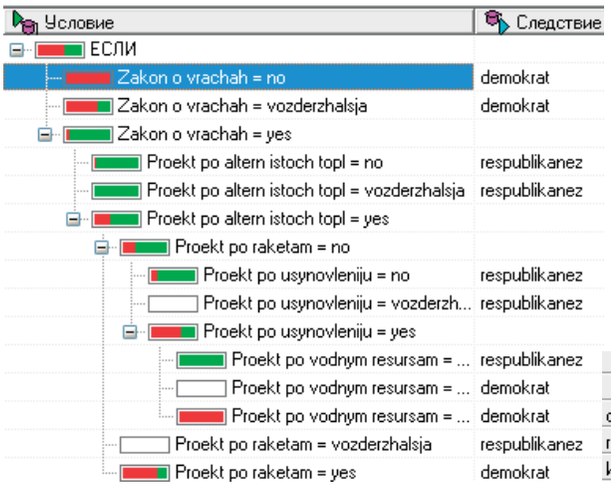
	Классифицировано			
Фактически	Iris-setosa	Iris-versicolor	Iris-virginica	Итого
Iris-setosa	50			50
Iris-versicolor		49	1	50
Iris-virginica		2	48	50
Итого	50	51	49	150

Рис. 2. Дерево решений и таблица сопряженности для «Ирисов Фишера» (Deductor)

```
Decision tree:

Zakon o vrachah in {no,vozderzhalsja}: demokrat (241/5)
Zakon o vrachah = yes:
:...Proekt po altern istoch topl in {no,
    :                               vozderzhalsja}: respublikanez (131/3)
    Proekt po altern istoch topl = yes:
    :...Proekt po raketam = vozderzhalsja: respublikanez (0)
        Proekt po raketam = yes: demokrat (5/1)
        Proekt po raketam = no:                                          (a)   (b)    <-classified as
        :...Proekt po usynovleniju = vozderzhalsja: respublikanez (0)   ----  ----
            Proekt po usynovleniju = yes: demokrat (6/2)                 143     8    (a): class respublikanez
            Proekt po usynovleniju = no: respublikanez (17/2)              5   244    (b): class demokrat
```

Рис. 3. Дерево решений и таблица сопряженности для «Голосования депутатов» (See5)



	Классифицировано		
Фактически	demokrat	respublikanez	Итого
demokrat	244	5	249
respublikanez	6	145	151
Итого	250	150	400

Рис. 4. Дерево решений и таблица сопряженности для «Голосования депутатов» (Deductor)

у обоих алгоритмов примерно одинаковая и составляет менее 1 сек.

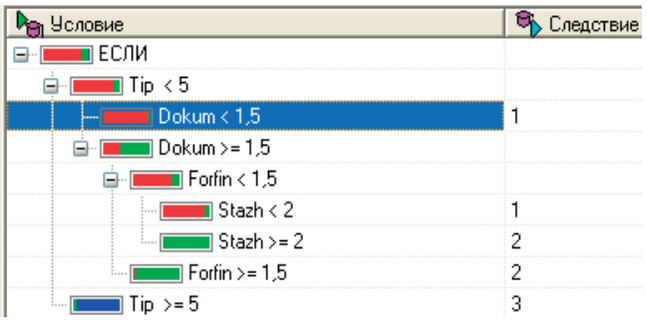
Таким образом, можно сделать вывод, что, как и заявлял Д. Куинлан, алгоритм C5.0 действительно строит более компактные деревья решений, чем его предшественник, а также обладает высокой скоростью построения классификационных моделей. Тем не менее достоверность результатов работы алгоритма C4.5 выше, чем у C5.0. В то же время нельзя исключать, что вышеуказанные ре-

```
Decision tree:

Dokum <= 1: 1 (303/6)
Dokum > 1:
:...Tip > 4: 3 (16/1)
    Tip <= 4:
    :...ForfinForfin <= 1: 1 (33/6)
        ForfinForfin > 1: 2 (48/2)

 (a)   (b)   (c)    <-classified as
----  ----  ----
 324     2          (a): class 1
  12    46     1    (b): class 2
               15    (c): class 3
```

Рис. 5. Дерево решений и таблица сопряженности для «Абитуриентов» (See5)



	Классифицировано			
Фактически	1	2	3	Итого
1	324	2		326
2	9	49	1	59
3			15	15
Итого	333	51	16	400

Рис. 6. Дерево решений и таблица сопряженности для «Абитуриентов» (Deductor)

Таблица 1. Количество правил

Алгоритм/Набор данных	Ирисы	Голосование	Абитуриенты
C4.5	5	11	5
C5.0	4	5	4

зультаты объясняются тем, что в их распоряжении имелась демонстрационная версия системы See5, которая может обрабатывать файлы, содержащие не более 400 записей.

Литература

1. Большие данные (Big Data) // URL: http:// www.tadviser.ru/index.php (д.о. 10.10.2015).
2. Data Mining – интеллектуальный анализ данных // URL: http://www.inftech. webservis .ru/it /database/datamining/ar2.html (д.о. 10.10.2015).
3. Топ-10 data mining-алгоритмов простым языком // URL: http://habrahabr.ru/company /itinvest/ blog/262155/ (д.о. 11.10.2015).
4. Деревья решений – общие принципы работы // URL: http://www.gotai.net/documents/doc-msc-006.aspx (д.о. 12.10.2015).
5. The ID3 Algorithm // URL: http://www.cise. ufl. edu/~ddd/cap6635/Fall-97/Short-papers/2. htm (д.о. 12.10.2015).
6. Сидоров А.В. Алгоритмы создания дерева принятия решений // URL: http://econf.rae. ru/ pdf/2014/03/3245.pdf (д.о. 13.10.2015).
7. Is See5/C5.0 Better Than C4.5? // URL: http:// rulequest.com/see5-comparison.html. (д.о. 15.10.2015).
8. Deductor – описание аналитической платформы // URL: http://bitconsulting. ru/product/olap/ (д.о. 17.10.2015).
9. Studio // URL: http://basegroup.ru/deductor/ components/studio (д.о. 17.10.2015).
10. Data Mining Tools See5 and C5.0 // URL: http:// rulequest.com/see5-info.html (д.о. 17.10.2015).
11. Iris Data Set // URL: http://archive. ics.uci.edu/ ml/datasets/Iris (д.о. 19.10.2015).

Поступило 20.10.2015

Пальмов Сергей Вадимович, к.т.н., доц. Кафедры информационных систем и технологий (ИСТ) Поволжского государственного университета телекоммуникаций и информатики (ПГУТИ). Тел.: 8-927-706-61-21. E-mail: psv@psuti.ru

Мифтахова Альфия Асхатовна, аспирант Кафедры ИСТ ПГУТИ. Тел. 8-937-795-01-95. E-mail: miftaxovaa@mail.ru

COMPARISON OF CLASSIFICATION ALGORITHMS C4.5 AND C5.0

Palmov S.V., Miftakhova A.A.
Povolzhskiy State University of Telecommunication and Informatics, Samara, Russian Federation
E-mail: psv@psuti.ru

This work compares features of tree decision algorithms C4.5 and C5.0, which are the most effective data mining classification tool. We considered two software tools: analytics platform Deductor and system See5. Three data sets were tested to improve comparative analysis accuracy. First is conventional Fisher's iris data set, second contains information about US Congress deputy votes (distribution Deductor), and third includes information about applicants of the one of Russian Federation universities. According to test results, C5.0 builds more compact decision trees, but its operation speed is almost the same to C4.5 under reducing of classification model validity. However, we do not preclude that these results can be explained by using of See5 system demo version that provides only files processing with no more 400 entries.

Keywords: Data Mining, decision tree, C4.5, C5.0, Deductor, See5

DOI: 10.18469/ikt.2015.13.4.18

Palmov Sergey Vadimovich, PhD in Technical Science, Assistant Professor of the Department of Information Systems and Technologies, Povolzhsky State University of Telecommunications and Informatics, Russian Federation, Samara. Tel. +79377950195. E-mail: psv@psuti.ru.
Miftakhova Alfiya Ashatovna, PhD-student, Department of the Information Systems and Technologies, Povolzhsky State University of Telecommunications and Informatics, Russian Federation, Samara. Tel.: +79377950195. E-mail: miftaxovaaa@mail.ru

References

1. Bolshie dannye (Big Data). Available at: http://www.tadviser.ru/index.php /Stat'ja:Bol'shie_dannye_(Big_Data) (accessed 10.10.2015).
2. Data Mining – intellektual'nyj analiz dannyh. Available at: http://www.inftech.webservis.ru/it/database/datamining/ar2.html (accessed 10.10.2015).
3. Top-10 data mining-algoritmov prostym jazykom. Available at: http://habrahabr.ru/company/itinvest/blog/262155/ (accessed 11.10.2015).
4. Derevja reshenij – obshhie principy raboty. Available at: http://www.gotai.net/documents/doc-msc-006.aspx (accessed 12.10.2015).
5. The ID3 Algorithm. Available at: http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm (accessed 12.10.2015).
6. Sidorov A.V. Algoritmy sozdanija dereva prinjatija reshenij. Available at: http://econf.rae.ru/pdf/2014/03/3245.pdf (accessed 13.10.2015).
7. Is See5/C5.0 Better Than C4.5? Available at: http://rulequest.com/see5-comparison.html (accessed 15.10.2015).
8. Deductor – opisanie analiticheskoj platform. Available at: http://bitconsulting.ru/product/olap/ (accessed 17.10.2015)
9. Studio. Available at: http://basegroup.ru/deductor/components/studio (accessed 17.10.2015).
10. Data Mining Tools See5 and C5.0. Available at: http://rulequest.com/see5-info.html (accessed. 17.10.2015).
11. Iris Data Set. Available at: http://archive.ics.uci.edu/ml/datasets/Iris (accessed 19.10.2015).

Received 20.10.2015