Моделирование алгоритма машинного обучения «Случайный лес» математическим аппаратом теории сетей Петри

Д.А. Петросов

Финансовый университет при Правительстве РФ, Москва

Аннотация: В статье рассматривается возможность моделирования алгоритма машинного обучения «случайный лес» с применением математического аппарата теории сетей Петри. Предложенный подход базируется на использовании трех видов расширений сетей Петри: классического, цветных сетей, а также вложенных сетей. Для этого в работе рассматриваются общая структура деревьев решений и правила построения моделей на основе двудольного направленного графа с последующим переходом на алгоритм машинного обучения «случайный лес». В статье приведены примеры моделирования данного алгоритма сетями Петри с формированием дерева достижимых маркировок, которое соответствует работе как деревьев решений, так и «случайному лесу».
Ключевые слова: сеть Петри, дерево решений, случайный лес, машинное обучение, теория сетей Петри, двудольный направленный граф, интеллектуальные системы, эволюционные алгоритмы, системы поддержки принятия решений, математическое моделирование, теория графов, имитационное моделирование, искусственные нейронные сети, вложенные сети Петри, цветные сети Петри.

Алгоритм машинного обучения «случайный лес» получил широкое распространение при решении большого класса задач:

– регрессионный анализ (например, использование алгоритма «случайный лес» в регрессии Надарая-Уотсона [1]);

– распознавание образов (например, распознавание распространения вредных растений [2]);

– прогнозирование (например, применение алгоритма «случайный лес» в задачах прогноза прибыли организации [3]);

– классификация (например, классификация классов грунта [4]) и т.д.

Популярность данного подхода связана с относительной простотой использования и высоким качеством решения задач. В некоторых современных исследованиях говорится о способности конкурирования данного алгоритма с искусственными нейронными сетями, то есть некоторый

класс задач, в котором использование данного алгоритма позволяет добиваться не меньшей результативности по соотношению с искусственными нейронными сетями.

В рамках данного исследования требуется рассмотреть возможность моделирования алгоритма машинного обучения «случайный лес» с использованием математического аппарата теории сетей Петри. Сети Петри получили широкое распространение при решении различного класса задач:

– проектирование архитектуры программных средств [5];

– моделирование работы серверных мощностей и рабочих станций [6];

– моделирование работы искусственных нейронных сетей [7, 8] и т.д.

Для этого предлагается рассмотреть три расширения двудольного направленного графа:

– классические сети (с применением ингибиторных дуг, позволяющим уточнять правила ветвления на узлы в дереве решений);

– цветные сети (для реализации возможности ветвления на узлы в дереве решений);

– вложенные сети Петри (для реализации дополнительной обработки меток в рамках работы переходов) [9].

Выбранные расширения позволяют моделировать работу деревьев решений, с учетом специфики создания и должны обеспечивать полноценную работу правил ветвления «ЕСЛИ - ТО».

Рассмотрим пример дерева решений, на рисунке 1 показана простая модель, позволяющая принять решение относительно страхования автомобиля [10].

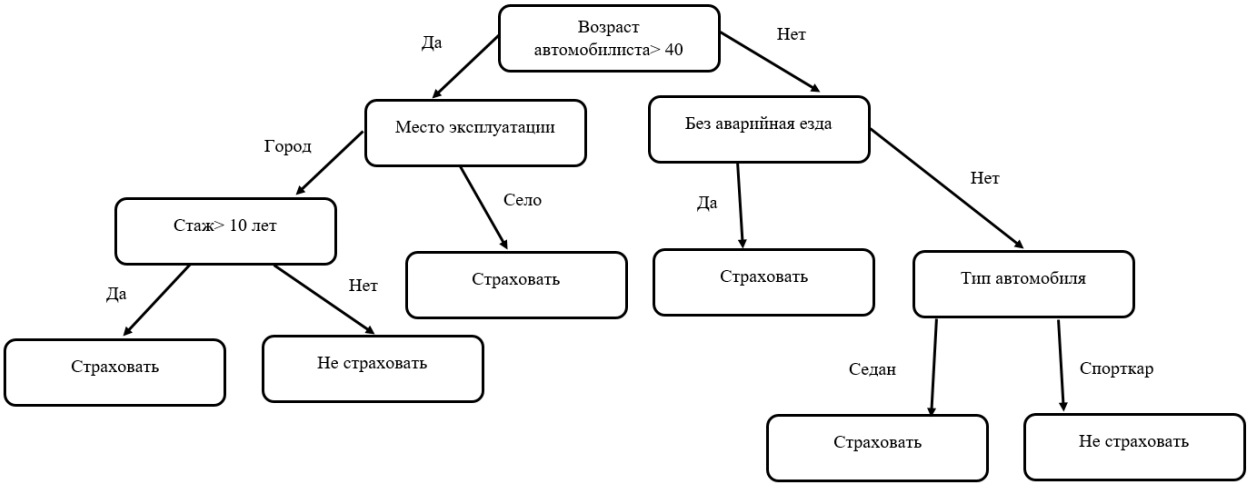

Рис.1. – Модель дерева решений для страхования автомобиля

Сеть Петри, моделирующую представленное дерево решений следует представить следующим образом (1)

$$PN_{TreeDes} = <P, T, L, M0>, \quad (1)$$

где

P – позиции;

T – переходы;

L- дуги;

M_0 – начальная маркировка.

В свою очередь позиции сети следует разделить следующим образом (2)

$$P = <Leaf, Data, Node>, \quad (2)$$

где

Leaf – листья дерева решений, при этом:

$$Leaf = \{Pleaf_0, \ldots, Pleaf_k\}, \quad (4)$$

где

k – количество позиций листьев дерева решений.

Data – данные с которыми работает дерево решений, при этом:

$$Data = \{PData_{,0}, \ldots, PData_{,n}\}, \quad (5)$$

где

n – количество позиций с данными для обработки.

Node – узлы дерева решений.

$$Node = \{PNode_{,0}, \ldots, PNode_{,i}\}, \quad (6)$$

где

j – количество позиций узлов.

В свою очередь при использовании в модели ингибиторных дуг:

$$L = <L, L_{ing}>, \quad (7)$$

где

L – обычное соединение позиций и переходов, при этом:

$$L = \{l_{1,c}, \ldots, l_{r,c}\}, \quad (8)$$

где

r – количество дуг;

c – цвет дуги.

L_{ing} – ингибиторное соединение (позволяющее обеспечить активацию переходов при отсутствии во входящей позиции метки), при этом:

$$L_{ing} = \{ling_{0,c} \ldots, ling_{o,c}\},. \quad (9)$$

где

o – количество ингибиторных дуг.

При использовании классического расширения переходы в сети Петри могут быть представлены следующим образом (10).

$$T = \{t_{0,\,c}, \ldots, t_{f,c}\}, \quad (10)$$

где

f – количество переходов в модели;

В случае использования вложенных сетей Петри любой из множества переходов T может быть представлен сетью PN, с обрабатывающие метки в соответствии с представленной моделью и цветом c, тогда:

$$T = \{PNT_{0,c}, \ldots, PNT_{f,c}\}, \quad (11)$$

Начальная маркировка M_0, расположение меток в начальном состоянии сети, может быть представлена в следующем виде:

$$M_0=\{MLeaf_{0,c},\dots MLeaf_{k,c},, MData_{0,c},\dots MData_{n,c}, MNode_{0,c},\dots MNode_{j,c}\}, \quad (12)$$

Тогда (1) можно представить в следующем виде (13):

$$PN_{TreeDes} = <\{Pleaf_0, \dots, Pleaf_k\}, \{PData_0, \dots, PData_n\},\{PNode_0, \dots, PNode_i\},$$
$$\{PN_0,\dots,PNT_f\},\{l_1,\dots,l_r\},\{ling_{0,c},\dots, lingt_c\},\{MLeaf_{0,c}\dots, MLeaf_{k,c}, MData_{0,c},\dots MData_{n,c},$$
$$MNode_{0,c},\dots MNode_{j,c}\}>, \quad (13)$$

В соответствии с представленной моделью дерева решений (см. рис. 1) и (13) построим модель сети Петри, полностью удовлетворяющую как требованиям выбранного математического инструментария, так и модели дерева решений. Результаты моделирования представлены на рисунке 2.

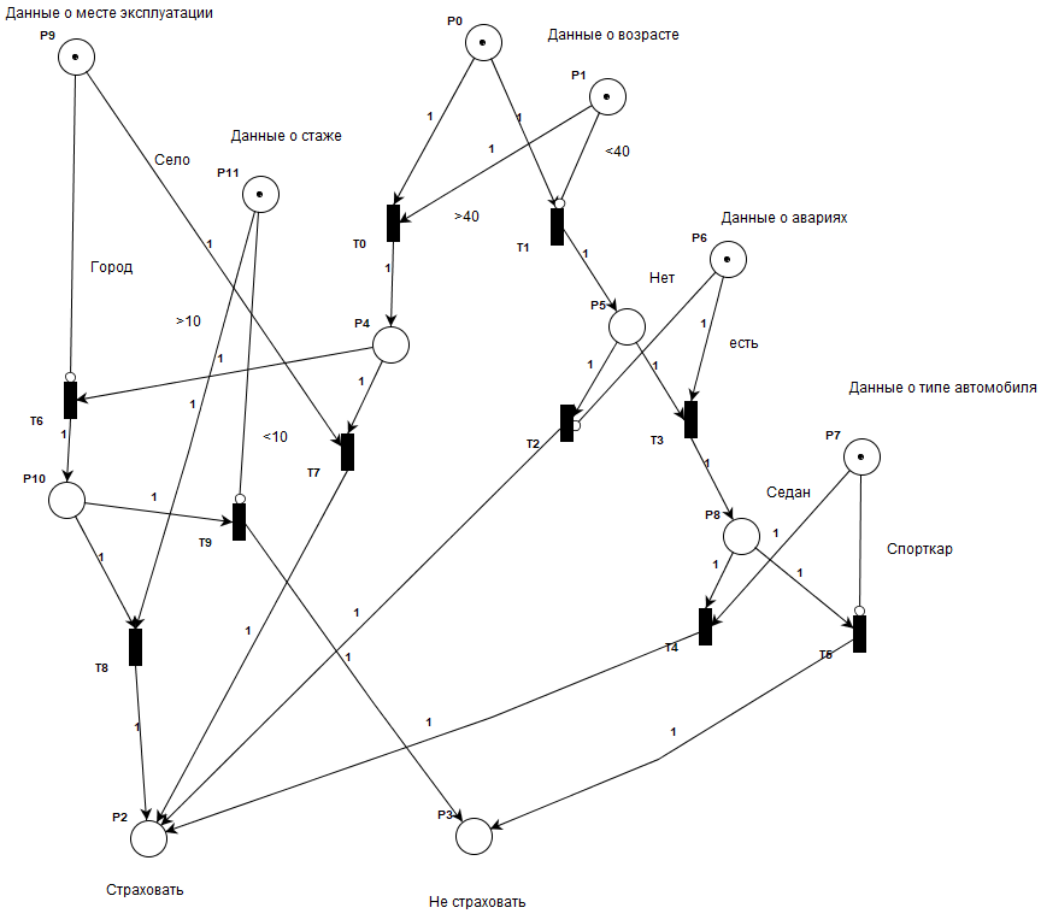

Рис.2. – Результат моделирования дерева принятия решений с использованием сети Петри

При решении поставленной задачи достаточно использовать классическое расширение сетей и дополнением в виде ингибиторных дуг, так как в модели дерева решений не используется более двух условий, соответственно в этом случае можно не вводить условия цветности меток, переходов и дуг, а достаточно проверки на наличие или отсутствие метки в соответствующей позиции.

Позиция *P0* является стартовой, то есть размещение метки в данной позиции запускает работу предложенной модели.

В предложенной модели позициями для хранения обрабатываемых данных являются:

$$Data = \{P1, P6, P7, P9, P11\}, \quad (14)$$

Данные позиции соединены с обычными и ингибиторными дугами с переходами, выполняющими работу по правилам «ЕСЛИ - ТО».

Для примера рассмотрим правило работы переходов *T0* и *T1*:

– правило работы *T0* – «ЕСЛИ в позиции *P0* и позиции *P1* есть хотя бы одна метка, ТО перемещаем метку в позицию *P4*»;

– правило работы *T1* – «ЕСЛИ в позиции *P0* есть метка, а в позиции *P1* нет метки, ТО перемещаем метку в позицию *P5*».

Соответственно при записи данных об автомобилисте в позиции P1 наличие метки говорит о том, что его возраст более 40 лет, а отсутствие метки говорит о том, что он моложе указанного возраста. На представленном рисунке, иллюстрирующем разработанную модель, даны пояснения работы переходов.

В качестве позиции *Node* определены:

$$Node = \{P4, P5, P8, P10\}, \quad (15)$$

Данные позиции используются в качестве узлов и для сохранения промежуточного результата работы сети.

В качестве позиций *Leaf* определены:

$$Leaf = \{P2, P3\}, (16)$$

В ходе вычислительных экспериментов были рассмотрены различные варианты начальных маркировок для класса позиций с данными. Результаты работы предложенной модели полностью соответствуют работе дерева решений. Таким образом можно говорить о том, что применение теории сетей Петри для моделирования деревьев решений вполне целесообразно с учетом деревьев различной сложности и глубины.

В соответствии с алгоритмом «случайный лес» данные и целевая переменная могут быть разделены для решения на разных деревьях решений и в задаче классификации окончательное решение принимается большим количеством голосов. Поэтому для моделирования алгоритма «случайный лес» с использованием сетей Петри целесообразно разработать модель голосования и выбора класса большинством деревьев.

В общем виде модель на основе сетей Петри будет выглядить следующим образом:

$$P=<PVote, PSave, PDesig>, (17)$$

где

PVote – позиции хранящие данные о решении (голосе) каждого дерева;

$$PVote = \{\{PVote_{1,1}, \ldots, PVote_{1,k}\}, \{PVote_{2,1}, \ldots, PVote_{2,k}\}, \ldots, \{PVote_{x,1}, \ldots, PVote_{x,k}\}\}, (16)$$

где

x – количество деревьев в «случайном лесу»;

$к$ – количество вариантов решения для каждого дерева.

PSave – позиции, хранящие данные о результате голосования;

$$PSave = \{PSave_1, PSave_2, \ldots, PSave_k\}, (17)$$

PDesig – позиции, определяющие решение алгоритма «случайный лес».

$$PDesign = \{PDesign_1, PDesign_2, \ldots, PDesign_k\}, (18)$$

Допустим, что в алгоритме «случайный лес» при решении задачи классификации автомобилистов опытным путем было определено, что оптимальное количество деревьев на представленных данных равно трем и каждое из представленных деревьев, на основе своей выборки данных, может принять решение или «Страховать», или «Не страховать».

Требуется разработать модель голосования, позволяющую путем выбора обычным большинством классифицировать автомобилиста и либо выдать страховой полис, либо отказать. На рисунке 3 показана предлагаемая модель голосования с использованием классического расширения сетей Петри.

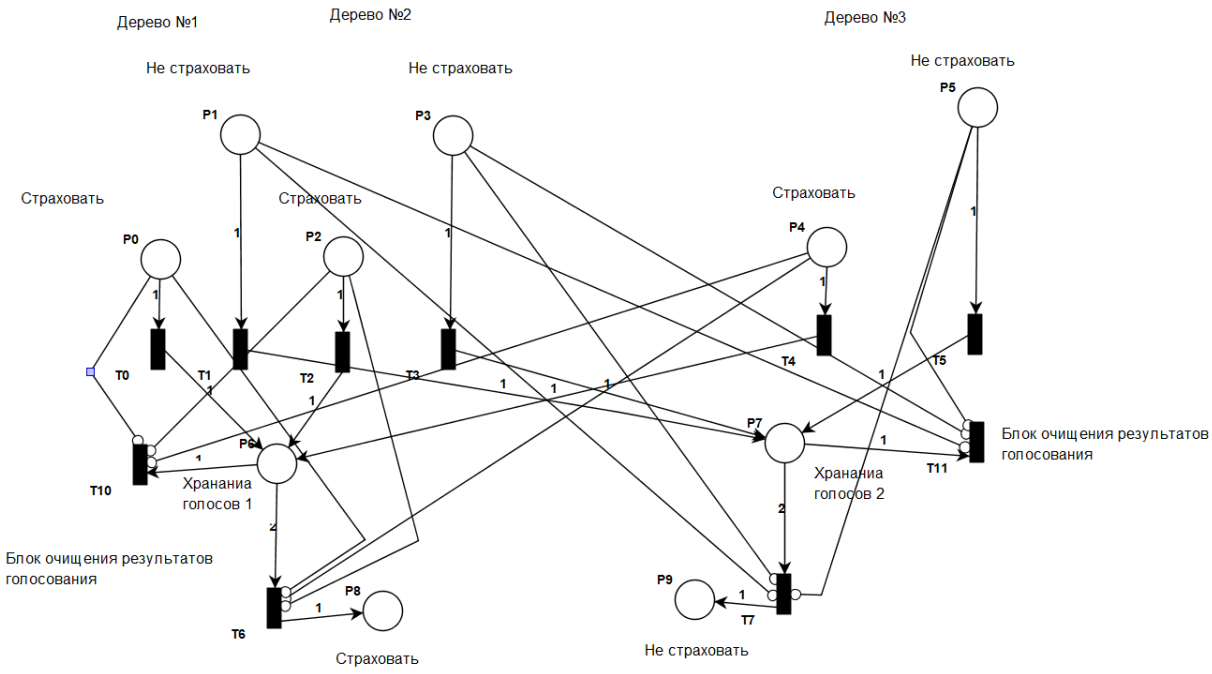

Рис. 3. – Пример модели голосования тремя деревьями

Выходы трех деревьев моделируются следующим образом

$$PVote = \{\{PVote_{1,1}, PVote_{1,2}\},\{PVote_{2,1}, PVote_{2,2}\},\{PVote_{3,1}, PVote_{3,2}\}\}, \quad (19)$$

$$PVote = \{\{P0, P1\},\{P2, P3\},\{P4, P5\}\}, \quad (20)$$

Хранилища голосов представлены следующим образом:

$$PSave = \{PSave_1, PSave_2\}, \quad (21)$$

$$PSave = \{P6, P7\}, \quad (22)$$

Результаты голосования моделируются следующим образом:

$$PDesign = \{PDesign_1, PDesign_2\}, \quad (23)$$

$$PDesign = \{P8, P9\}, \quad (24)$$

Для организации накоплений в хранилища голосов требуется организовать переход меток в соответствующие позиции с этой целью в модели введены следующие переходы:

– для учета варианта «Страховать»: *T0*, *T2*, *T4*, соответственно позиция *P6*;

– для учета варианта «Не страховать»: *T1*, *T3*, *T5*, соответственно позиция *P7*.

Так как количество деревьев равно трем, а решение принимается простым большинством, то, соответственно, достаточно соединить позиции хранения голосов с позициями решения через переходы (*T6*, *T7*) с дугой весом два в позиции *P8* и *P9*. При этом потребуется дополнительно установить ограничение на работу перечисленных переходов в виде ингибиторных дуг, связывающих их с соответствующими позициями результатов принятия решений деревьев. Это требуется для того, чтобы исключить преждевременное срабатывание переходов *T6* и *T7* до того момента, пока все деревья не отдадут свои голоса.

После отработки событий, связанных с голосованием, требуется выполнить очищение модели от оставшихся меток, для этого используются переходы *T10* и *T11*.

Таким образом задача моделирования как деревьев решений, так и алгоритма «случайный лес» полностью решена, как в общем виде, так и на рассмотренном примере.

Предложенный в рамках исследования подход для моделирования алгоритма «случайный лес» с использованием сетей Петри позволит дополнить модели генетических алгоритмов, описанных с применением

данного математического аппарата. Применение данного метода машинного обучения может быть использовано для решения задачи управления процессом поиска решений на основе эволюционной процедуры. Надстройка в виде модели сети Петри, описывающей работы алгоритма «случайный лес» может позволить выполнить классификацию состояния популяции и выполнить управляющее воздействие на операторы генетического алгоритма. С учетом того, что искусственные нейронные сети и алгоритм «случайный лес» могут решать задачи классификации с достаточно большой точностью, для решения задачи повышения эффективности работы генетического алгоритма в дальнейшем потребуется провести исследование по эффективности применения как подхода на основе искусственных нейронных сетей так и на основе алгоритма «случайного леса». Кроме этого следует отметить, что применение матричного подхода для описания процедуры классификации может позволить повысить быстродействие программных средств с применением технологии GPGPU, так как позволит распределить вычисления на множество вычислителей минуя работу блока ветвлений.

Благодарности. *Работа выполнена при финансовой поддержке РНФ (проект №23-31-00127)*

Литература

1. Utkin L., Konstantinov A. Random survival forests incorporated by the nadaraya-watson regression // Informatics and Automation. 2022. V. 21. № 5. pp. 851-880.

2. Yifter T.T., Razoumny Yu.N., Orlovsky A.V., Lobanov V.K. Monitoring the spread of sosnowskyi's hogweed using a random forest machine learning algorithm in google earth engine // Computer Research and Modeling. 2022. V. 14. № 6. pp. 1357-1370.

3. Ломакин Н.И., Марамыгин М.С., Положенцев А.А., Шабанов Н.Т., Наумова С.А., Старовойтов М.К. Модель глубокого обучения RF «случайный лес» для прогнозирования прибыли организации в условиях цифровой экономики //Международная экономика. 2023. № 11. С. 824-839.

4. Гущина О.А., Коржов А.С. Применение алгоритма случайного леса для автоматизации классификации категорий грунтов// Огарёв-Online. 2023. № 16 (201) URL: journal.mrsu.ru/arts/primenenie-algoritma-sluchajnogo-lesa-dlya-avtomatizacii-klassifikacii-kategorij-gruntov. (дата обращения: 24.09.2024)

5. Харахинов В.А., Сосинская С.С. Использование сетей Петри при проектировании архитектуры программного продукта для анализа данных с помощью нейронных сетей // Научный вестник Новосибирского государственного технического университета. 2018. № 4 (73). С. 91-100.

6. Тронин В.Г., Стецко А.А. Моделирование сервера и рабочей станции вычислительной сети с помощью раскрашенных сетей Петри //Программные продукты и системы. 2008. № 3. С. 95-97.

7. Петросов Д.А. Моделирование искусственных нейронных сетей с использованием математического аппарата теории сетей Петри //Перспективы науки. 2020. № 12 (135). С. 92-95.

8. Петросов Д.А. Кодирование маркировки сети Петри, моделирующей работу искусственной нейронной сети //Вопросы устойчивого развития общества. 2020. № 9. С. 433-438.

9. Ермакова В.О., Ломазова И.А. Трансляция вложенных сетей Петри в классические сети Петри для верификации разверток //Труды Института системного программирования РАН. 2016. Т. 28. №4. С. 115-136.

10. Loginom. Деревья решений: общие принципы, URL:loginom.ru/blog/decision-tree-p1 (дата обращения: 24.09.2024)

References

1. Utkin L.; Konstantinov A. Informatics and Automation 2022, V. 21, № 5, p. 851-880.

2. Yifter T.T.; Razoumny Yu.N.; Orlovsky A.V.; Lobanov V.K. Computer Research and Modeling 2022, V. 14, 6, p. 1357-1370

3. Lomakin N.I., Maramygin M.S., Polozhencev A.A., SHabanov N.T., Naumova S.A., Starovojtov M.K. Mezhdunarodnaya ekonomika 2023, 11. p. 824-839.

4. Gushchina O.A., Korzhov A.S. Ogaryov-Online 2023, 16 (201). URL: journal.mrsu.ru/arts/primenenie-algoritma-sluchajnogo-lesa-dlya-avtomatizacii-klassifikacii-kategorij-gruntov. (date assessed: 24.09.2024)

5. Harahinov V.A., Sosinskaya S.S. Nauchnyj vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta 2018, 4 (73). pp. 91-100.

6. Tronin V.G., Stecko A.A. Programmnye produkty i sistemy 2008, 3. pp. 95-97.

7. Petrosov D.A. Perspektivy nauki 2020, 12 (135), pp. 92-95.

8. Petrosov D.A. Voprosy ustojchivogo razvitiya obshchestva 2020, 9, p. 433-438.

9. Ermakova V.O., Lomazova I.A. Trudy Instituta sistemnogo programmirovaniya RAN 2016, T. 28, 4, pp. 115-136.

10. Loginom. Derev'ya reshenij: obshchie principy [Loginom. Decision trees: general principles]. URL:loginom.ru/blog/decision-tree-p1 (date assessed: 24.09.2024).

Дата поступления: 14.09.2024

Дата публикации: 20.10.2024