ВЕСТНИК НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО ЯДЕРНОГО УНИВЕРСИТЕТА «МИФИ», 2022, т. 11, № 6, с. 342–349

ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА

УДК 004.62

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ДЕРЕВЬЕВ РЕШЕНИЙ И НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧЕ КЛАССИФИКАЦИИ КРЕДИТНЫХ ОРГАНИЗАЦИЙ

© 2022 Е.П. Акишина[1], В.В. Иванов[1,2], А.В. Крянев[2*], А.С. Приказчикова[2]
[1]Объединенный институт ядерных исследований, Дубна, 141980, Россия
[2]Национальный исследовательский ядерный университет «МИФИ», Москва, 115409, Россия
**e-mail: AVKryanev@mephi.ru*
Поступила в редакцию: 12.01.2023
После доработки: 12.01.2023
Принята к публикации: 24.01.2023

В последние годы деревья решений и нейронные сети широко применяются в задачах компьютерного зрения, таких как распознавание объектов, классификация текстов, распознавание жестов, обнаружение спама, семантическая сегментация и кластеризация данных. В статье рассматривается применение методов деревьев решений и искусственных нейронных сетей в задаче классификации кредитных организаций как объектов экономической безопасности. Представлены результаты анализа данных о деятельности кредитных организаций с использованием разных методов деревьев решений: C5, CHAID, C&R и QUEST, а также нейронных сетей. Наивысшая общая точность классификации анализируемых объектов была достигнута с помощью алгоритма деревьев решений C5 и составила 81 %. Общая точность классификации при применении алгоритма CHAID составила 68 %, алгоритма C&R – 71 %, алгоритма QUEST – 66 %. На основании результатов алгоритма C5 сгенерирован набор правил для определения принадлежности банка к определенному классу. Согласно методам деревьев решений и нейронным сетям были отобраны наиболее информативные показатели деятельности кредитных организаций с точки зрения их разбиения на два класса: благонадежные и высоко-рисковые.

Ключевые слова: машинное обучение, деревья решений, нейронная сеть, классификация, кредитные организации, отмывание преступных доходов.
DOI: https://doi.org/10.26583/vestnik.2022.12

ВВЕДЕНИЕ

Как известно, деревья решений применяются в задачах классификации (принятие решения о принадлежности объекта к одному из классов) и регрессии (предсказание значения из непрерывного диапазона). Классификация и регрессия на основе деревьев решений используются в задачах распознавания текста, информационного поиска, распознавания речи, анализе изображений, обнаружении спама, распознавания жестов и др.

В деревьях решений используется автоматическая настройка параметров алгоритма на основе обучающей выборки. Деревья решений состоят из вершин, в которых записаны проверяемые условия (в нашем случае – показатели, принимающие те или иные значения), и листьев, в которых записаны «ответы» дерева. Обучение состоит в настройке условий в узлах дерева и ответов в его листьях с целью достижения максимальной точности классификации. Деревья решений позволяют производить сегментацию анализируемых объектов, осуществлять предсказание путем формирования логических правил, сокращать данные и идентифицировать взаимосвязи внутри классов [1].

На рис. 1, в качестве примера, представлено дерево решений для классификации кредитных организаций, построенное с использованием алгоритма CHAID [1]. Особенностями деревьев решений являются автоматический отбор признаков, интерпретируемость, управляемость обучения, зависимость от числа обучающих примеров разных классов, риск переобучения, достаточно большая обучающая выборка [2].

ПРИМЕНЕНИЕ НЕЙРОННОЙ СЕТИ К ИССЛЕДУЕМЫМ ДАННЫМ

В предыдущей работе [3] анализировалась банковская сфера в России, а именно, финансовая деятельность кредитных организаций в целях идентификации незаконной активности отдельных ее звеньев. Признаковое пространство, характеризующее операционную деятельность кредитных организаций, исследовалось посредством механизма искусственных нейронных

сетей. С помощью пакета программных продуктов Statistica 6 [4] была построена архитектура нейронной сети, использовались данные финансовой отчетности № 101. Выборка данных организаций была разбита на подвыборки – обучающая 50 % экз., контрольная – 25 % экз., тестовая – 25 % экз. По каждой выборке в программе был рассчитан показатель «Производительность». Отметим, что обучающая выборка используется для обучения модели, тестовая выборка – для оценки качества модели, контрольная выборка – для выбора наилучшей модели из имеющихся. Обучающая выборка представляла собой совокупность 23 непрерывных показателей и одного целевого категориального показателя – «Отзыв». Однако общую производительность по всей совокупности данных Statistica не формирует в отдельный показатель.

Перейдем к рассмотрению применения искусственных нейронных сетей к исследуемым данным о кредитных организациях. Искусственная нейронная сеть – упрощенная модель биологической нейронной сети, представляющая собой совокупность искусственных нейронов, взаимодействующих между собой [5]. В качестве модели нейронной сети использовался многослойный перцептрон, архитектура которого представлена на рис. 2.

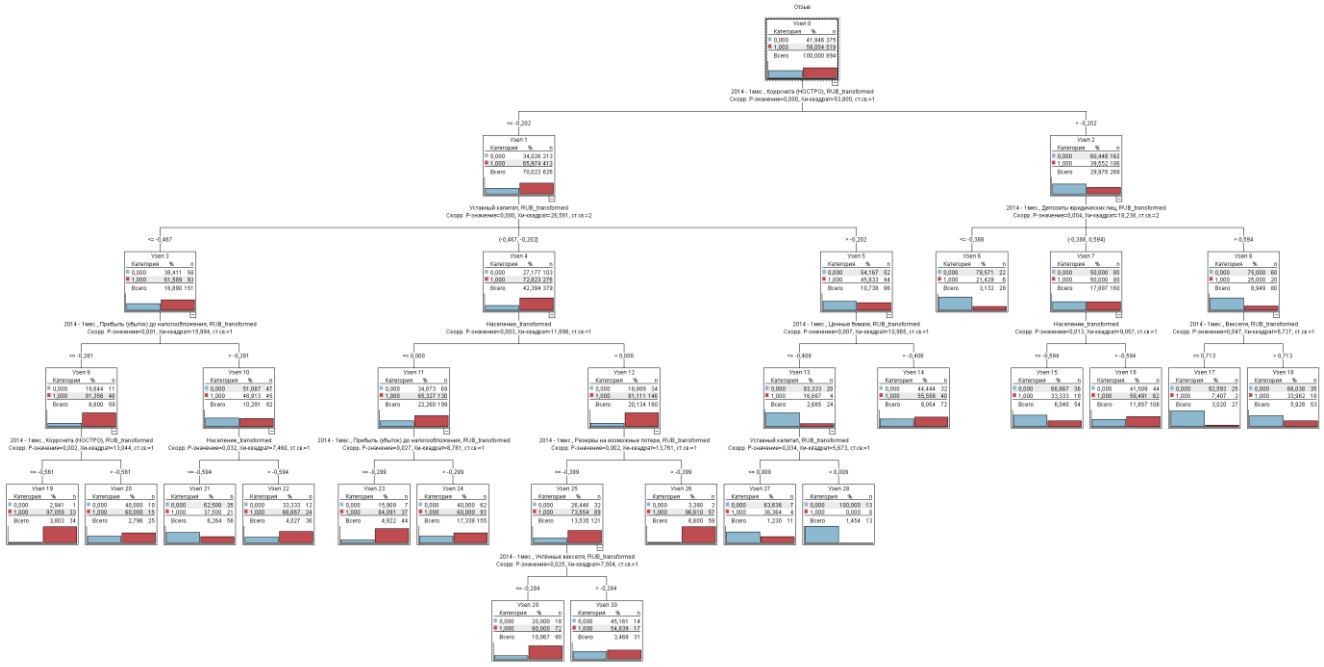

Рис. 1. Дерево решений для классификации кредитных организаций (алгоритм CHAID)

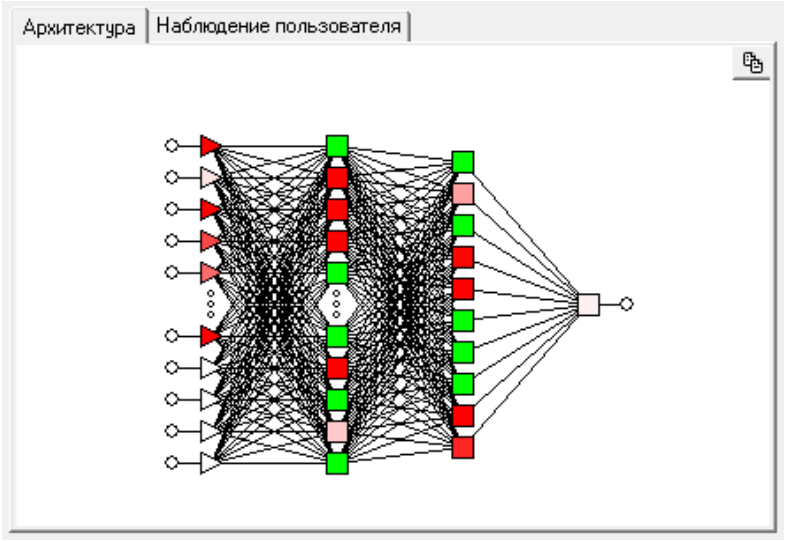

Рис. 2. Архитектура нейронной сети. Перцептрон

Наилучшей будет считаться модель нейронной сети, имеющая наивысшее значение показателя «Производительность». Производительность нейронной сети рассчитывается как отношение правильно классифицированных нейронной сетью объектов к общему количеству объектов:

$$\text{Производительность} = \frac{C}{G}, \qquad (1)$$

где C – количество правильно классифицированных объектов, а G – общее количество объектов в выборке.

ПРИМЕНЕНИЕ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ ФАКТОРНОГО АНАЛИЗА

Следующим этапом исследования явилось проведение анализа признакового пространства деятельности кредитных организаций с помощью метода главных компонент факторного анализа. В основе метода главных компонент лежат всевозможные линейные преобразования исходных признаков. В результате отработки метода из исходных признаков было сгенерировано 23 главных компоненты. На рис. 3 представлен график «каменистой осыпи», который наглядно демонстрирует вклад каждой главной компоненты в совокупную дисперсию. Из рисунка видно, что наибольший вклад дает первая главная компонента, а начиная с 12 компоненты вклад незначителен. Поэтому для дальнейшей работы можно сфокусироваться на 12 главных компонентах (из сгенерированных 23), общая совокупность дисперсии которых составляет 99 %. Авторами были проанализированы значения производительностей нейронных сетей, построенных на 23 и 12 компонентах.

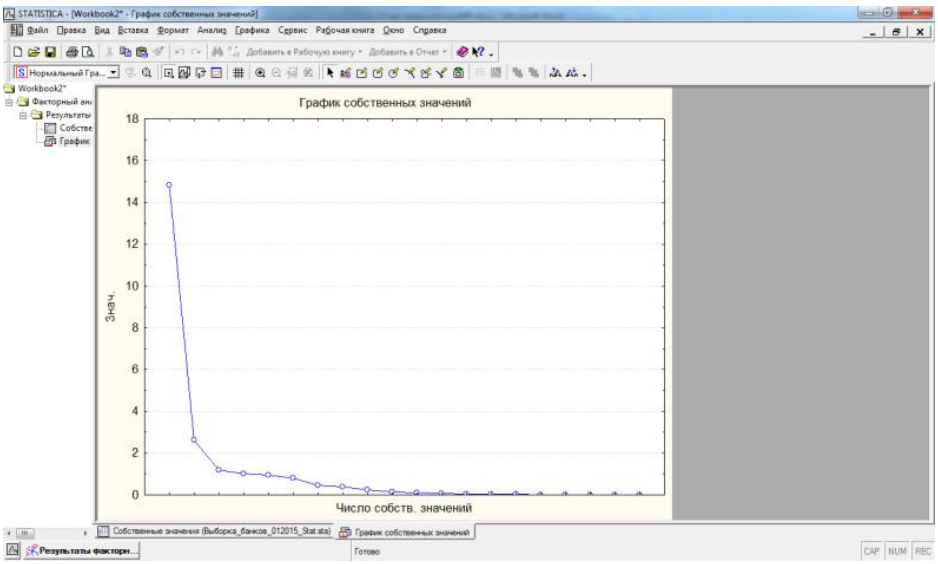

Рис. 3. График «каменистой осыпи»

На рис. 4 представлены результаты корреляционного анализа исходных показателей и сгенерированных главных компонент. Интерпретируем первые три главные компоненты. Первый фактор имеет сильную связь (высокое значение коэффициента корреляции) со следующими исходными показателями: чистые активы, ценные бумаги, кредиты, кредиты организациям, кредиты физическим лицам, основные средства, средства клиентов, средства организаций на расчетных счетах (р/с), депозиты юридических лиц, вклады физических лиц, резервы на возможные потери, капитал. В связи с чем данный фактор характеризует финансовую состоятельность банка, его платежеспособность и активную работу с клиентами, как с физическими, так и с юридическими лицами. Второй фактор имеет высокое значение коэффициента корреляции с уставным капиталом, кредитами другим банкам, облигациями. Таким образом, вторая главная компонента характеризует финансовую возможность банка предоставлять в долг денежные средства, а также гарантировать интересы его кредиторов. Третий фактор имеет высокое значение корреляции со счетами в Банке России и коррсчетами (НОСТРО). Третья главная компонента характеризует состояние корреспондентских счетов банка, открытых в Центральном Банке России или в других банках для осуществления взаимных расчетов.

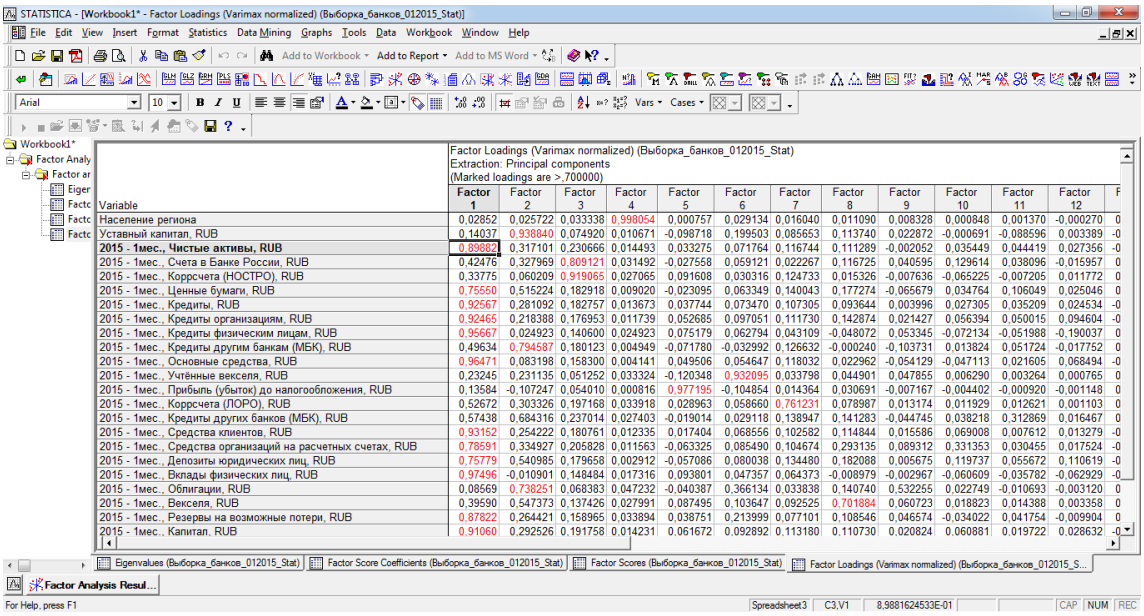

Рис. 4. Корреляция исходных показателей и главных компонент

РАСЧЕТ ТОЧНОСТИ КЛАССИФИКАЦИИ

При сравнении методов деревьев решений и нейронной сети необходимо учесть следующее. Пакет Statistica не формирует единый показатель производительности по всей совокупности данных. А эффективность применения деревьев характеризуется показателем точности классификации. В связи с чем для процедуры сравнения данных методов следует привести показатели классификации к единой шкале измерения. Рассчитаем общую точность классификации *Accuracy* объектов:

$$Accuracy = \frac{K}{G} \qquad (2)$$

(K – количество правильно классифицированных объектов, а G – общее количество объектов в выборке [6]) с использованием нейронной сети для данных 2014 и 2015 гг.

$$Accurancy_{2014} = \frac{623}{894} = 0,7, \qquad (3)$$

$$Accurancy_{2015} = \frac{579}{814} = 0,71. \qquad (4)$$

Аналогичные вычисления были проведены и для расчета общей точности классификации объектов с использованием нейронной сети на усеченной выборке, содержащей 12 главных компонент. Так, для данных 2014 г. значение точности классификации объектов нейронной сети на усеченной выборке составило 63 %, а для 2015 г. – 66 %.

У метрики *Accuracy* есть особенность, она подразумевает у всех объектов одинаковый вес, что может быть не корректно в случае неравномерного распределения объектов выборки по классам. И тогда у классификатора больше информации по одному классу и, соответственно, меньше – по другому. А значит, в рамках большего класса принятые решения могут быть более адекватные. Чтобы избежать этой проблемы, можно провести сравнение методов по точности внутри каждого класса. Точность классификации в пределах класса – это доля объектов, действительно, принадлежащих данному классу, относительно всех объектов, которые система отнесла к этому классу. Эти значения рассчитываются на основании матрицы ошибок (confusion matrix). Пусть дана выборка x_i ($i = 1, \ldots, N$, y_i – метка класса i-го объекта, $y_i \in \{1, \ldots, C\}$), каждый объект которой относится к одному из C классов, и классификатор a, который эти классы предсказывает. Матрицей ошибок для такого классификатора называется следующая матрица (5):

$$M = \left\{m_{ij}\right\}_{i,j=0}^{C}, \qquad m_{ij} = \sum_{k=0}^{N} \left\| \left[a\left(x_k\right) = j\right] \right\| \left[y_k = i\right]. \qquad (5)$$

Такая матрица показывает, сколько объектов класса i были распознаны как объекты класса j. Эта информация позволяет понять не только, сколько ошибок делает алгоритм, но и то, насколько он точен. В случае бинарной классификации метка класса y принимает значение 0 (положительный класс) или 1 (отрицательный). Вводятся четыре величины:

- истинно положительные объекты (*TP – true positive*), которые были классифицированы как положительные и действительно являются положительными (принадлежащими к данному классу);
- истинно отрицательные объекты (*TN – true negative*), которые были классифицированы как отрицательные и действительно отрицательные (принадлежащими к данному классу);
- ложноположительные объекты (*FP – false positive*), которые были классифицированы как положительные, но фактически отрицательные;
- ложноотрицательные объекты (*FN – false negative*), которые были классифицированы как отрицательные, но фактически положительные.

Величины соответствуют элементам матрицы ошибок:

$$TP = \sum_{i=0}^{n} \left[a\left(x_i\right) = +1 \right] \left[y_i = +1 \right], \quad (6)$$

$$TN = \sum_{i=0}^{n} \left[a\left(x_i\right) = -1 \right] \left[y_i = -1 \right], \quad (7)$$

$$FP = \sum_{i=0}^{n} \left[a\left(x_i\right) = +1 \right] \left[y_i = -1 \right], \quad (8)$$

$$FN = \sum_{i=0}^{n} \left[a\left(x_i\right) = -1 \right] \left[y_i = +1 \right]. \quad (9)$$

Общий вид матрицы ошибок представлен в табл. 1.

Таблица 1. Матрица ошибок

Data		Экспертная оценка	
		Положительная	Отрицательная
Оценка классификатора	Положительная	*TP*	*FN*
	Отрицательная	*FP*	*TN*

Составим аналогичную матрицу ошибок по классу благонадежных организаций по алгоритму C5 за 2014 и 2015 гг. (табл. 2, 3).

Таблица 2. Матрица ошибок 2014

2014	Экспертная оценка	
Оценка классификатора	221	154
	18	501

Таблица 3. Матрица ошибок 2015

2015	Экспертная оценка	
Оценка классификатора	222	132
	44	416

Имея такую матрицу, точность для каждого класса *Precision* рассчитывается отдельно по формуле

$$Preision = \frac{TP}{TP + FP}, \quad (10)$$

соответственно, точность для класса – отношение количества правильно отнесенных объектов по классу ко всем объектам класса в выборке [6].

ПРИМЕНЕНИЕ ДЕРЕВЬЕВ РЕШЕНИЙ

На завершающем этапе исследования было реализовано построение моделей деревьев решений с использованием программного продукта IBM SPSS Modeler [5]. Для построения деревьев решений существуют разные алгоритмы, в том числе C5, CHAID, C&R и QUEST. Эти алгоритмы реализуются рекурсивно, подгруппы разбиваются на все меньшие и меньшие блоки до тех пор, пока дерево не будет завершено. В результате отбора лучших моделей для анализируемого набора данных установлено, что самая высокая общая точность классификации на использованных данных достигнута с использованием алгоритма C5 и составила 81 %. Общая точность классификации при применении алгоритма CHAID – 68 %, алгоритма C&R – 71 %, алгоритма QUEST – 66 %.

После построения деревьев решений с использованием разных алгоритмов, важно оценить полученные результаты. Для этого можно использовать такие способы, как:

1) сегментация (позволяет идентифицировать тех, кто входит в определенную группу, а также вероятности попадания конкретного объекта в группу);

2) предсказание (позволяет сформировать правила и использовать их для предсказания будущих событий);

3) сокращение данных и экранирование переменных (позволяет выбрать полезное подмножество предикторов из большого набора переменных для использования при построении формальной параметрической модели);

4) идентификация взаимодействия (позволяет установить взаимосвязи, которые принадлежат определенной подгруппе, и указать их в формальной параметрической модели).

Рассмотрим отдельные результаты применения алгоритмов деревьев решений. Согласно

алгоритму C5 наиболее важными показателями финансовой деятельности, с точки зрения классификации кредитных организаций на благонадежные и высоко-рисковые, являются кредиты, корсчета (НОСТРО) и ценные бумаги (правая часть на рис. 5).

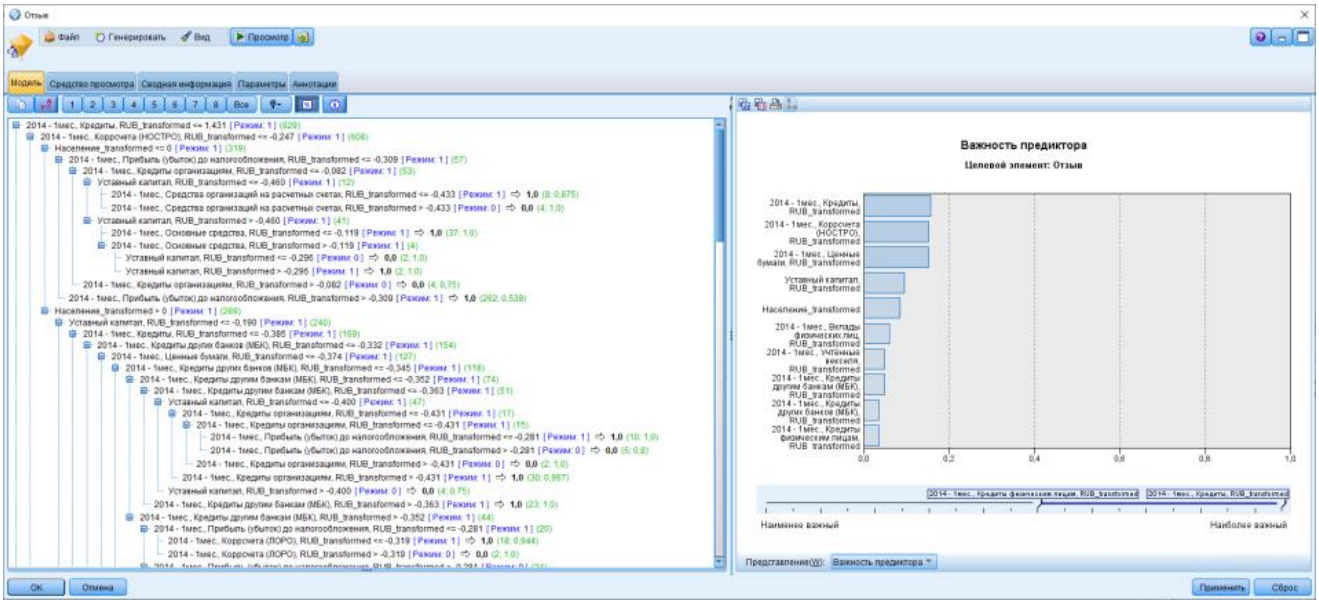

Рис. 5. Результаты классификации кредитных организаций с использованием алгоритма C5 деревьев решений

На рис. 5 (левая часть) представлен набор правил для алгоритма C5. Наборы правил получаются из дерева решений и в некотором смысле представляют собой упрощенную или очищенную версию информации, найденную деревом решений.

Проанализировав визуальную структуру деревьев решений, построенных с использованием разных алгоритмов, отметим, что дерево решений C5 обладает самой сложной структурой. Количество уровней в дереве составляет 16 ед. Для сравнения, в дереве решений CHAID – 5 уровней, в дереве решений C&R – 5 уровней, в дереве решений QUEST – 4 уровня.

Ниже приведена сравнительная табл. 4 точностей классификации (*Accuracy*) нейронной сети и рассмотренных методов деревьев решений, а также точностей для классов благонадежных и неблагонадежных организаций (*Precision*) на данных стандартной финансовой отчетности № 101 за 2014 и 2015 гг.

Таблица 4. Сравнительная таблица значений точности классификации

Метод	Точность, все организации 2014	Точность, все организации 2015	Точность, благонадежные 2014	Точность, благонадежные 2015	Точность, ненадежные 2014	Точность, ненадежные 2015
НС (исходная выборка)	70 %	71 %	63 %	65 %	77 %	77 %
НС (усеченная выборка – 12 ГК)	63 %	66 %	63 %	66 %	63 %	65 %
Алгоритм C5	81 %	78 %	92 %	83 %	76 %	76 %
Алгоритм CHAID	68 %	68 %	62 %	61 %	74 %	76 %
Алгоритм C&R	70 %	71 %	62 %	64 %	76 %	78 %
Алгоритм QUEST	66 %	65 %	64 %	66 %	67 %	64 %

АНАЛИЗ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ И ВЫВОДЫ

В заключение можно сделать вывод, что максимальную общую точность классификации рассматриваемых объектов показал алгоритм C5 деревьев решений. Далее приведена табл. 5 наиболее значимых показателей, согласно модели нейронной сети и разновидностям построенных деревьев решений. В данном исследовании наиболее информативными показателями при проведении классификаций кредитных организаций на благонадежные и высокорисковые являются «Коррсчета (НОСТРО)», «Уставный капитал», «Ценные бумаги» и др.

Таблица 5. Наиболее информативные показатели НС и деревьев решений

Нейронная сеть	Алгоритм C5	Алгоритм CHAID	Алгоритм C&R	Алгоритм QUEST
Уставный капитал	Кредиты	Корсчета (НОСТРО)	Корсчета (НОСТРО)	Корсчета (НОСТРО)
Корсчета (НОСТРО)	Корсчета (НОСТРО)	Уставный капитал	Прибыль (убыток) до налогообложения	Население региона
Средства организаций на р/с	Ценные бумаги	Население региона	Средства организаций на р/с	Капитал
Ценные бумаги	Уставный капитал		Уставный капитал	
Вклады физических лиц	Население региона		Население региона	

Итак, в настоящей работе рассмотрено применение искусственных нейронных сетей и методов деревьев решений в задаче классификации кредитных организаций как объектов экономической безопасности. Представлены результаты анализа данных о деятельности кредитных организаций с использованием разных методов деревьев решений: C5, CHAID, C&R и QUEST, а также нейронных сетей. Наивысшая общая точность классификации анализируемых объектов была достигнута с помощью алгоритма деревьев решений C5 и составила 81 %. На основании результатов данного алгоритма сгенерирован набор правил для определения принадлежности банка к определенному классу. Согласно методам деревьев решений и нейронным сетям были отобраны наиболее информативные показатели деятельности кредитных организаций с точки зрения их разбиения на два класса: благонадежные и высокорисковые.

При внедрении в структуру Единой информационной системы Росфинмониторинга рассмотренных выше методов деревьев решений и искусственных нейронных сетей аналитикам службы станет возможным осуществлять оперативную автоматизированную классификацию кредитных организаций с целью установления факта их вовлеченности в противоправную деятельность по легализации денежных средств, полученных преступным путем.

СПИСОК ЛИТЕРАТУРЫ

1. *Паклин Н.Б., Орешков В.И.* Бизнес-аналитика: от данных к знаниям. учеб. пособие. 2-е изд. СПб.: Питер, 2013. С. 428–472. ISBN 978-5-459-00717-6.

2. *Левитин А.В.* Алгоритмы: введение в разработку и анализ. М.: Вильямс, 2006. С. 409–417. 576 с. ISBN 978-5-8459-0987-9.

3. *Иванов В.В., Акишина Е.П., Приказчикова А.С.* Применение нейронных сетей и метода главных компонент для идентификации кредитных организаций, потенциально вовлеченных в процесс по легализации преступных доходов // Известия Иссык-Кульского форума бухгалтеров и аудиторов стран Центральной Азии. 2022 № 2(37). С. 294–296.

4. Электронный учебник StatSoft. [Электронный ресурс]. URL: Statisticahttp://statsoft.ru (дата обращения 10.12.2022).

5. *Haykin.* Simon Neural networks and learning machines / Simon Haykin. 3rd ed. p. cm. Rev. ed of: Neural networks. 2nd ed., 1999.

6. *Гладилин П.Е., Боченина К.О.* Технологии машинного обучения. СПб: Университет ИТМО, 2020. 75 с.

7. IBM SPSS Modeler. [Электронный ресурс]. URL: https://www.ibm.com/products/spss-modeler (дата обращения 10.12.2022).

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ДЕРЕВЬЕВ РЕШЕНИЙ И НЕЙРОННЫХ СЕТЕЙ
В ЗАДАЧЕ КЛАССИФИКАЦИИ КРЕДИТНЫХ ОРГАНИЗАЦИЙ

Vestnik Natsional'nogo issledovatel'skogo yadernogo universiteta «MIFI», 2022, vol. 11, no. 6, pp. 342–449

THE COMPARATIVE ANALYSIS OF DECISION TREES AND NEURAL NETWORKS METHODS IN THE CREDIT INSTITUTIONS CLASSIFICATION PROBLEM

E.P. Akishina[1], V.V. Ivanov[1,2], A.V.Kryanev[2*], A.S. Prikazchikova[2]
[1]Joint Institute for Nuclear Research, Dubna, 141980 Russia
[2]National Research Nuclear University ««MEPhI»«, Moscow, 115409, Russia
**e-mail: AVKryanev@mephi.ru*

Received January 12, 2023; revised January 12, 2023; accepted January 24, 2023

In recent years, decision trees and neural network have been widely used in computer vision problems such as object recognition, text classification, gesture recognition, spam detection, semantic segmentation and data clustering. The article discusses the decision tree and neural networks methods application in the problem of classifying credit institutions as economic security objects. The analysis results of data on credit institutions activities of using different methods of decision trees: C5, CHAID, C&R and QUEST, as well as neural networks are presented. The highest overall classification accuracy of the analyzed objects was achieved using the C5 decision tree algorithm and amounted to 81 %. The overall classification accuracy using the CHAID algorithm was 68 %, the C&R algorithm was 71 %, and the QUEST algorithm was 66 %. Based on the C5 algorithm results, a set of rules was generated to determine whether a bank belongs to a certain class. According to the methods of decision trees and neural networks, the most informative performance indicators of credit institutions were selected in terms of their division into two classes: trustworthy and high-risk.

Keywords: machine learning, decision trees, neural network, classification, credit institutions, money laundering.

REFERENCES

1. *Paklin N.B., Oreshkov V.I.* Biznes-analitika: ot dannyh k znaniyam. ucheb. posobie. 2-e izd. [Business Analytics: From Data to Knowledge, textbook allowance. 2nd ed.]. St. Petersburg. Piter Publ., 2013. P. 428–472. ISBN 978-5-459-00717-6.

2. *Levitin A.V.* Algoritmy: vvedenie v razrabotku i analiz. [Algorithms. Introduction to development and analysis]. M.: Williams Publ., 2006. P. 409–417. ISBN 978-5-8459-0987-9.

3. *Ivanov V.V., Akishina E.P., Prikazchikova A.S.* Primenenie nejronnyh setej i metoda glavnyh komponent dlya identifikacii kreditnyh organizacij, potencial»no vovlechennyh v process po legalizacii prestupnyh dohodov [Application of neural networks and the method of principal components to identify credit institutions potentially involved in the process of money laundering]. Izvestiya Issyk-Kul»skogo foruma buhgalterov i auditorov stran Central»noj Azii, 2022. № 2 (37). P. 294–296.

4. Elektronnyj uchebnik StatSoft [Electronic textbook StatSof]. Available at: http://statsoft.ru (accessed 10.12.2022).

5. *Haykin*. Simon Neural networks and learning machines. Simon Haykin. 3rd ed. p. cm. Rev. ed of: Neural networks. 2nd ed., 1999.

6. *Gladilin P.E., Bochenina K.O*. Tekhnologii mashinnogo obucheniya [Machine learning technologies]. St. Petersburg: ITMO University Publ., 2020. 75 p.

7. IBM SPSS Modeler. Available at: https://www.ibm.com/products/spss-modeler (accessed 10.12.2022).