УДК 519.17
doi:10.21685/2227-8486-2021-1-3

АЛГОРИТМ БИНАРНОЙ КЛАССИФИКАЦИИ НА ОСНОВЕ ГРАФОВ ПРИНЯТИЯ РЕШЕНИЙ В ЗАДАЧАХ КРЕДИТНОГО СКОРИНГА

А. Н. Кисляков

Владимирский филиал Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации, г. Владимир, Россия
ankislyakov@mail.ru

Аннотация. *Актуальность и цели.* Рассмотрена актуальная проблема построения графов принятия решений оптимальной структуры, которые используются для решения задач бинарной классификации и создания прогностических моделей социально-экономических показателей. Цель работы заключается в обобщении опыта построения деревьев и графов принятия решений и исследовании качества классификационных моделей на их основе. *Материалы и методы.* Показаны примеры реализации алгоритмов на основе деревьев и рандомизированных ансамблей ориентированных ациклических графов принятия решений (DAG или джунгли решений) для задачи кредитного скоринга как одно из направлений модификации ансамблевых алгоритмов на основе деревьев решений. Основным отличием графов принятия решений от деревьев принятия решений является наличие узлов дерева, которые являются бинарными классификаторами, могут быть связаны с другими узлами, иерархически не связанными с родительским узлом. Таким образом, двоичный граф по сравнению с деревом решений может иметь не только корневые, но и расщепленные узлы, а также листья. Поскольку признаки, соответствующие разным классам, имеют различные значения показателя, корневой узел может разделить их в соответствии с этой особенностью, также это могут сделать и дочерние узлы. *Результаты.* На основе показателя энтропии дерева может быть рассчитан информационный прирост, который позволит оптимизировать структуру графа путем минимизации общей взвешенной энтропии набора значений предикторов. В итоге вырабатываются не только правила разделения, но и правила связи признаков внутри дерева, поэтому при меньшей глубине построения граф решений имеет большую способность описания взаимосвязей пространства признаков сложной системы. *Выводы.* Описана возможность преодоления проблемы неустойчивости конечных предсказаний моделей на основе деревьев решений на относительно небольших выборках данных путем перехода к графовым моделям классификации. Исследования показали, что графы принятия решений являются наиболее эффективным алгоритмом классификации, который показывает наилучшие результаты на небольшой обучающей выборке.

Ключевые слова: деревья решений, графы решений, бинарная классификация, машинное обучение

Для цитирования: Кисляков А. Н. Алгоритм бинарной классификации на основе графов принятия решений в задачах кредитного скоринга // Модели, системы, сети в экономике, технике, природе и обществе. 2021. № 1. С. 29–41. doi:10.21685/2227-8486-2021-1-3

THE ALGORITHM FOR BINARY CLASSIFICATION ON GRAPH-BASED DECISION-MAKING IN THE TASKS OF CREDIT SCORING

A.N. Kislyakov

Vladimir branch of the Russian Academy of National Economy and Public Administration under the President of the Russian Federation, Vladimir, Russia
ankislyakov@mail.ru

Abstract. *Background.* The work is devoted to the actual problem of constructing decision graphs of an optimal structure, which are used to solve problems of binary classification and create predictive models of socio-economic indicators. The aim of the work is to generalize the experience of constructing decision trees and graphs and to study the quality of classification models based on them. *Materials and methods.* Examples of implementation of algorithms based on trees and randomized ensembles of oriented acyclic decision graphs (DAG or jungle of decisions) for the problem of credit scoring as one of the directions of modification of ensemble algorithms based on decision trees are shown. The main difference between decision graphs and decision trees is the presence of tree nodes that are binary classifiers and can be connected to other nodes that are not hierarchically connected to the parent node. Thus, a binary graph compared to a decision tree can have not only root nodes, but also split nodes, as well as leaves. *Results.* Since the features corresponding to different classes have different values of the indicator, the root node can divide them according to this feature, as well as the child nodes can do it. Based on the entropy index of the tree, an information gain can be calculated, which will allow optimizing the graph structure by minimizing the total weighted entropy of a set of predictor values. As a result, not only the separation rules are developed, but also the rules for connecting features within the tree, so with a lower depth of construction of the decision graph, it has a greater ability to describe the relationships of the feature space of a complex system. *Conclusions.* The possibility of overcoming the problem of instability of finite predictions of models based on decision trees on relatively small data samples by switching to graph classification models are describe. Studies have shown that decision graphs are the most efficient classification algorithm that shows the best results on a small training sample.

Keywords: decision trees, decision graphs, binary classification, machine learning

For citation: Kislyakov A.N. The algorithm for binary classification on graph-based decision-making in the tasks of credit scoring. *Modeli, sistemy, seti v ekonomike, tekhnike, prirode i obshchestve = Models, systems, networks in economics, technology, nature and society*. 2021;1:29–41. (In Russ.). doi:10.21685/2227-8486-2021-1-3

Введение

Деревья принятия решений являются одним из самых распространенных и универсальных алгоритмов машинного обучения и применяются как для задач классификации, так и для задач регрессии [1]. На основе деревьев решений построено огромное количество предиктивных моделей, в том числе и для задачи оценки кредитоспособности клиентов коммерческих банков [2].

Алгоритмы на основе деревьев решений позволяют выявить потенциально возможные закономерности и связи между отдельными компонентами социально-экономической системы и предсказать новые факты путем оценки значения целевого признака *y* (отклика) для любого объекта по его описанию $X = (x_1, x_2, ..., x_n)$ – набору независимых переменных, называемых предикторами [1, 3]. При создании прогностических моделей основной задачей явля-

ется предсказание величины целевого признака y на основании наблюдаемой вариации значений переменных $x_1, x_2, ..., x_n$ без исследования структуры внутренних взаимосвязей между переменными и/или сравнительной оценки силы их влияния на отклик.

Следует отметить, что при реализации проектов по анализу данных с применением алгоритмов машинного обучения с учителем, к которым относятся деревья решений, особое внимание следует обратить на качество набора данных, применяемого в процессе обучения модели.

В большинстве ситуаций моделируемые процессы, а следовательно, и описывающие их наборы данных имеют различное количество предикторов при ограниченном объеме выборки, что вызывает сложности при обучении классификационных алгоритмов.

В этой связи актуальной задачей является разработка подходов к исследованию социально-экономических систем, которые повышают точность и адекватность прогностических моделей в задачах классификации. Разработка подобных высоко интерпретируемых моделей может быть осуществлена с применением алгоритмов на основе деревьев решений, основным преимуществом которых является гибкость и интерпретируемость результатов анализа. Также важным преимуществом деревьев решений является небольшая вычислительная нагрузка при работе с большими объемами данных и признаков, высокая устойчивость к выбросам и возможность применения в задачах уменьшения размерности. Кроме того, одно из самых полезных свойств деревьев решений состоит в возможности наглядного отображения процесса обучения модели, что позволяет найти взаимосвязь между предикторами.

Цель работы заключается в обобщении опыта построения деревьев и графов принятия решений и исследовании эффективности и качества классификационных моделей на их основе применительно к задачам бинарной классификации при ограниченном объеме выборки.

Методы исследования

Деревья решений являются иерархической структурой, в которой каждый внутренний узел выполняет оценку признака с помощью правила классификации, каждая ветвь представляет результат классификации, а каждый лист (терминальный узел) содержит метку класса [4]. Такое построение позволяет гибко и эволюционно обучать модель на основе деревьев. Однако подобные модели могут оказаться существенно неустойчивыми. Небольшие изменения в обучающей выборке данных могут привести к существенным изменениям в структуре дерева и в итоге повлиять на качество конечных предсказаний.

Важной настройкой алгоритма является такой параметр, как глубина обучения – количество узлов дерева, используемых для классификации по одному из признаков [5]. По причине того, что алгоритмы на основе деревьев решений относятся к «жадным», они имеют склонность к переобучению модели, что значительно усложняет оптимизацию деревьев. Снижение эффекта переобучения может быть достигнуто путем управления сложностью модели, например, с помощью различных критериев остановки, таких как ограничение глубины построения дерева и последующая «обрезка» (pruning) отдель-

ных ветвей деревьев, что в итоге накладывает неявные ограничения на возможности алгоритма классификации, делая его не столь простым и универсальным [6].

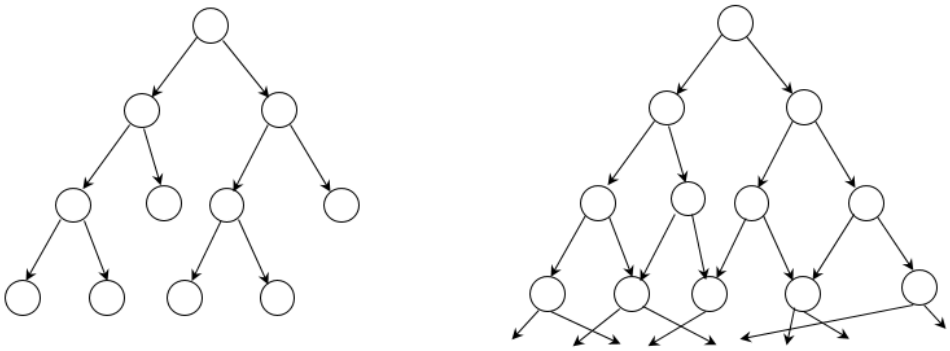
На качество обучения, конечно же, влияет и объем выборки, который требуется для построения точной классификационной модели. Однако если на текущем наборе данных не удается выстроить качественную модель, то необходимы модификации базового алгоритма деревьев решений. В основе создания оптимальных правил классификации на выборке, которая не позволяет построить модель с приемлемым уровнем точности, лежат более сложные ансамблевые методы, такие как, например, random forest (случайный лес) [7, 8] с последующей выборкой данных для обучения модели (bagging) или комбинированием работы нескольких алгоритмов на основе мажоритарного или мягкого голосования (boosting). Ансамблевые методы также накладывают ряд ограничений на исходный набор данных (датасет), размер которого должен быть достаточно большим, чтобы охватить большую часть сложности базового распределения, чтобы выборка из датасета была хорошим приближением к выборке из реального распределения (репрезентативность). Кроме того, выборки должны быть слабо коррелированными (независимыми).

В целом ансамблевые методы делают процесс принятия решений более устойчивым к отсутствующим данным и несовершенству базового варианта построения дерева. Однако исследования показывают, что для наборов данных с относительно большим количеством предикторов и объемом выборки, где глубокие деревья более точны, чем ансамблевые методы, наблюдается экспоненциальный рост количества узлов дерева, который создает проблему для реализации этого алгоритма на ограниченных аппаратных ресурсах [9, 10].

Другой подход к повышению эффективности алгоритмов на основе деревьев решений представляет собой возможность использования деревьев с многомерным откликом [1]. Этот подход также является эффективным лишь при наличии большого количества точек выборки при небольшом количестве признаков, иначе модель становится сложной и трудно интерпретируемой и не дает существенного прироста точности по сравнению с классическим алгоритмом построения дерева решений. Не следует забывать, что любое статистическое усреднение либо упрощение модели негативно сказывается на интерпретируемости результатов работы, поэтому для достижения лучшего результата на относительно небольшой по объему выборке необходимо всестороннее изучение пространства признаков объектов системы.

Одним из направлений модификации ансамблевых алгоритмов на основе деревьев решений является использование рандомизированных ансамблей ориентированных ациклических графов принятия решений (decision acyclic graph DAG) [11, 12] в качестве средства получения компактных и в то же время точных классификационных моделей. В русскоязычных источниках этот подход называется графом решений, в англоязычной литературе чаще встречается наименование «джунгли» принятия решений [9] по аналогии с популярным алгоритмом бэггинга на деревьях принятия решений – random forest (случайного леса).

Основным отличием джунглей решений от деревьев принятия решений является наличие узлов дерева, которые хоть и являются по-прежнему бинарными, но могут быть соединены с другими узлами, иерархически не связанными с родительским узлом (рис. 1).



а) *б*)

Рис. 1. Логика построения дерева решений (*а*) и графа решений (*б*)

Двоичный граф по сравнению с деревом решений может иметь не только корневые, но и расщепленные узлы, а также листья. Несмотря на то, что алгоритм построения графа принятия решений также является «жадным», в нем есть существенные отличия от алгоритма на основе деревьев решений: на начальном этапе также выбирается лист и разбивается на два листа с помощью узла решения – правила бинарной классификации, т.е. если существует необходимость разделить объекты по классам: добросовестный заемщик/не добросовестный заемщик, то для обучения графа будет использоваться маркированный набор признаков. Поскольку признаки, соответствующие разным классам, имеют различные значения показателя, корневой узел может разделить их в соответствии с этой особенностью, это могут сделать и дочерние узлы. Это приводит к появлению одинаковых значений признака в разных поддеревьях. Однако если возникает ситуация, что два таких узла связаны с одинаковыми распределениями классов, то, объединив их, можно получить один узел с обучающими примерами из обоих типов данного признака. Это помогает уловить степень изменчивости, присущую обучающим данным, и уменьшить сложность классификатора. В итоге вырабатываются не только правила разделения, но и правила связи признаков внутри дерева, поэтому при меньшей глубине построения дерева граф-джунгли решений имеет большую способность описания взаимосвязей пространства признаков системы.

Рассмотрим множество узлов на двух последовательных уровнях принятия решения DAG (см. рис. 1,*б*). Граф решений состоит из набора родительских N_p и дочерних N_c узлов. Глубина построения графа может быть обозначена как $M = |N_c|$. Пусть θ_i – это правило классификации для родительского узла $i \in N_p$, а S_i – набор значений предикторов (x, y) для узла i. Зная θ_i и S_i, можно вычислить набор значений предикторов узла i, проходящих через его левую и правую ветви соответственно [5, 9]:

$$S_i^L(\theta_i) = \{(x,y) \in S_i \mid f(\theta_i, x) \le 0\}; \; S_i^R(\theta_i) = S_i \setminus S_i^L(\theta_i). \qquad (1)$$

Если $l_i \in N_c$ – значение левого внешнего ребра от родительского узла $i \in N_p$ к дочернему узлу и $r_i \in N_c$ – значение правого внешнего ребра, тогда

набор значений предикторов, которые достигают любого родительского узла $j \in N_c$, вычисляется как

$$S_j\left(\{\theta_i\},\{l_i\},\{r_i\}\right)=\left[\bigcup_{i\in N_p,\, l_i=j} S_i^L(\theta_i)\right]\cup\left[\bigcup_{i\in N_p,\, r_i=j} S_i^R(\theta_i)\right]. \qquad (2)$$

Целевая функция *E*, оптимизирующая глубину построения DAG, является функцией от S_j. Таким образом, можно сформулировать задачу обучения графа решения оптимальной структуры как задачу минимизации целевой функции *E* в зависимости от количества операций разделения, а также операций левого и правого объединения. Целевая функция оптимизации количества узлов DAG может быть записана в виде

$$\min_{\{\theta_i\},\{l_i\},\{r_i\}} E\left(\{\theta_i\},\{l_i\},\{r_i\}\right). \qquad (3)$$

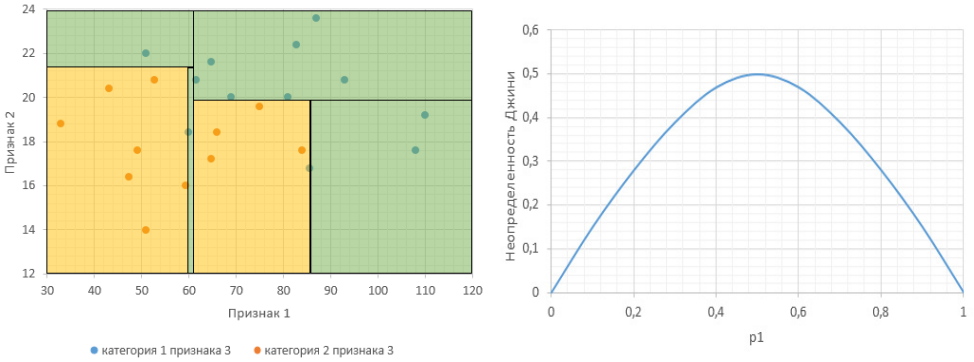
Таким образом, задача классификации на основе DAG формируется как задача минимизации энергии графа – суммы абсолютных величин собственных значений матрицы смежности графа. На практике используют два метода ограничения глубины оптимального DAG [6]. Первый метод основан на оптимизации количества узлов DAG, а второй – на оптимизации количества ветвей, исходящих из родительских узлов. Из полученных решений выбирается модель, которая будет иметь минимальный показатель *E*. Разбиение листа и объединение двух листьев не влияет на остальную структуру дерева. В этой связи основным преимуществом метода является возможность более качественной классификации на ограниченном наборе данных.

В алгоритмах классификации для выбора структуры правил и критериев бинарной классификации (чаще в моделях, основанных на алгоритмах деревьев решений) используется показатель Gini impurity. Не имея устоявшегося названия в русском языке, он трактуется как «примесь», «неоднородность Джини» или «неопределенность Джини» и характеризуется долей «примеси» точек из других классов в текущем варианте разбиения, т.е. описывает «чистоту» или однородность выделенных классов [6, 13]. Данный показатель называется неопределенностью Джини, так как он связан с показателем информационной энтропии и выражается как

$$I\left(A_k\right)=1-\sum_{k=1}^{m} p_k^2, \qquad (4)$$

где *m* – количество классов целевой переменной при номерах классов $k = 1, 2, \ldots, m$. где p_k –доля точек в прямоугольнике *A* (рис. 2,*а*), принадлежащих классу *k*. Эта мера принимает значения от 0 (когда все точки принадлежат к одному классу) и $(m - 1)/m$ (когда все *m* классов представлены одинаковым количеством точек).

Значения индекса Джини в зависимости от p_k для двух классов показаны на рис. 2. Мера примеси находится на своем пике, когда $p_k = 0{,}5$ (т.е. когда прямоугольник содержит 50 % точек другого класса). Этот показатель также может быть использован для оценки качества классификации модели на основе графа принятия решений.



а) *б*)

Рис. 2. Пример разбиения пространства признаков (*а*) и интерпретация неопределенности Джини (*б*) в задачах построения моделей классификации

Еще одним показателем оптимальности построения классификационной модели является энтропия Шеннона [14, 15], которая рассчитывается на основе гистограммы состояний и признаков процесса. Основной проблемой интерпретации показателя энтропии является «зашумленность» наблюдаемого процесса случайными событиями. Но при этом энтропия применяется для оценки «примесей» в задачах классификации и для прямоугольника A_k (см. рис. 2) определяется следующим образом:

$$H\left(A_k\right) = -\sum_{k=1}^{m} p_k \cdot \log_2\left(p_k\right). \tag{5}$$

Этот показатель колеблется от 0 (наиболее чистое разделение или все точки принадлежат к одному классу) до $\log_2(m)$ (в прямоугольнике все m классов представлены одинаково). В случае двух классов мера энтропии достигает максимума при $p_k = 0{,}5$.

На основе показателя энтропии дерева может быть рассчитан показатель information gain, который в русскоязычной формулировке звучит как информационный прирост, или усиление информативности, или взаимная информация [5]. В рамках решаемой задачи оптимизации структуры графа требуется минимизация общей взвешенной энтропии набора значений предикторов, определяемой как

$$E\left(\{\theta_i\},\{l_i\},\{r_i\}\right) = \sum_{j \in N_c} \left|S_j\right| \cdot H\left(S_j\right), \tag{6}$$

где S_j определяется по формуле (1), а $H\left(S_j\right)$ – энтропия Шеннона меток класса y в обучающем наборе предикторов $(x, y) \in S$. Чем больше информационный прирост, тем полезнее взаимная информация всех узлов графа и тем он более полно описывает пространство признаков при минимальном количестве узлов. В случае, когда количество дочерних узлов M строго равно удвоенному количеству родительских узлов, т.е. $M = 2N_p$, граф вырождается в дерево решений [9].

Результаты и обсуждение

В качестве набора данных был взят стандартный набор, используемый для обучения моделей задач кредитного скоринга – разновидности задач бинарной классификации. Модель должна дать ответ на вопрос, с какой долей вероятности клиент способен просрочить выплату кредита более чем на 90 дней. Датасет состоит более чем из 300 тыс. записей и находится в открытом доступе по ссылке https://www.kaggle.com/brycecf/give-me-some-credit-dataset. Описание полей набора данных представлено в табл. 1.

Таблица 1

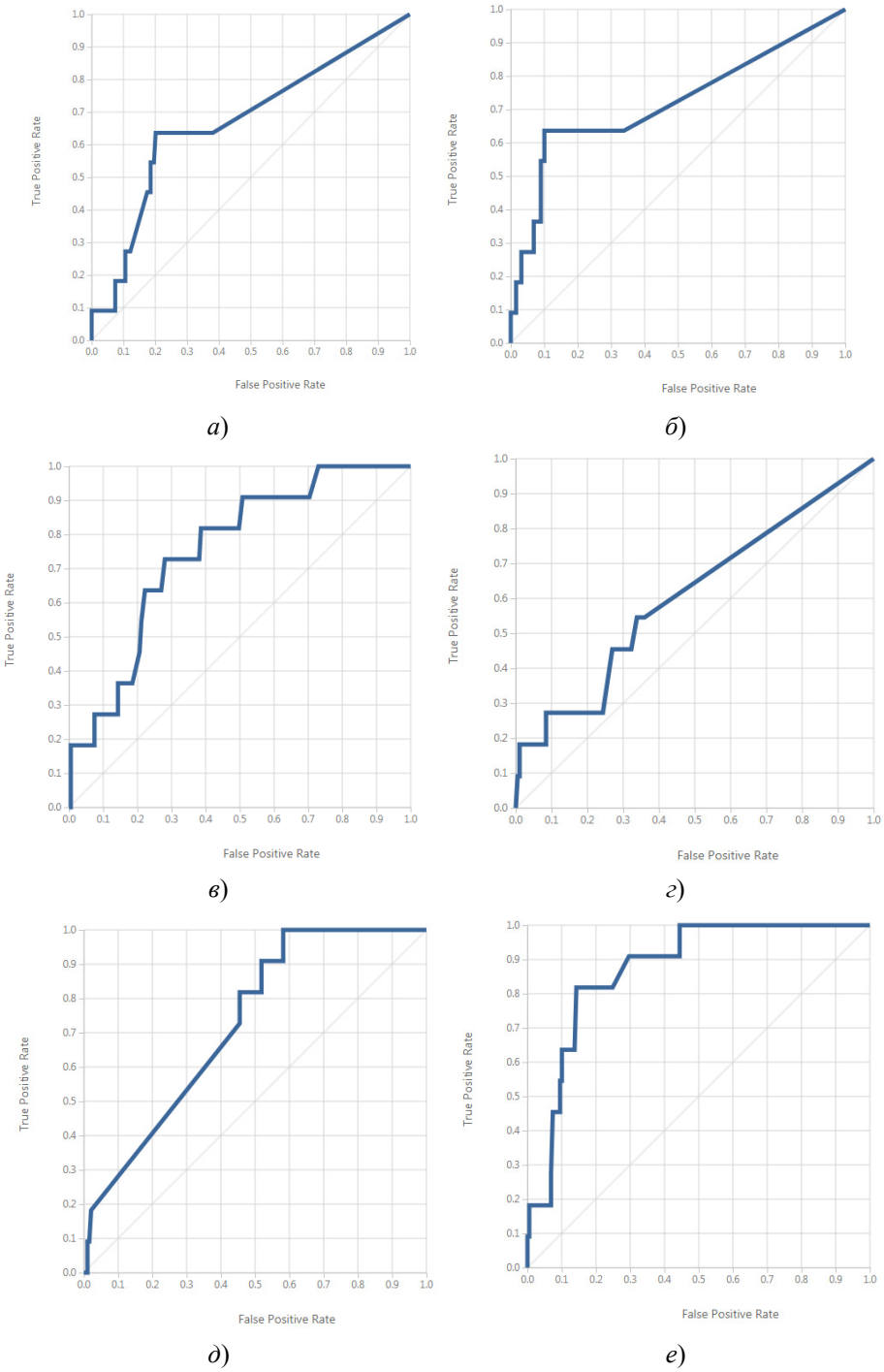
Описание полей набора данных

Имя переменной	Описание	Тип
SeriousDlqin2yrs	Человек имеет свыше 90 дней просроченной задолженности за последние 2 года	Бинарный
RevolvingUtilization OfUnsecuredLines	Общий баланс по кредитным картам и личным кредитам за исключением недвижимости и без рассрочки долга	Процентный
Age	Возраст заемщика в годах	Целое
NumberOfTime30-59DaysPastDueNotWorse	Количество случаев, когда заемщик просрочил выплату на 30–59 дней за последние 2 года	Целое
DebtRatio	Ежемесячные платежи по долгам, алименты, расходы на проживание, приведенные к ежемесячному валовому доходу	Процентный
MonthlyIncome	Ежемесячный доход	Вещественный
NumberOfOpen CreditLinesAndLoans	Количество открытых кредитов (рассрочка, например, автокредит или ипотека) и кредитных линий (кредитные карты)	Целое
NumberOfTimes90DaysLate	Количество просроченных платежей заемщика на 90 дней или более	Целое
NumberRealEstate LoansOrLines	Количество ипотечных кредитов и кредитов на недвижимость, включая кредитные линии на покупку жилья	Целое
NumberOfTime60-89DaysPastDueNotWorse	Количество случаев, когда заемщик просрочил выплату на 60–89 дней за последние 2 года	Целое
NumberOfDependents	Количество иждивенцев в семье заемщика, исключая его самого (т.е. супруга, дети и т.д.)	Целое

Моделирование выполнялось средствами открытой low-code платформы Microsoft Azure Machine Learning Studio и инструментария языка программирования R. Для того чтобы показать преимущество использования графов принятия решений перед прочими классификационными алгоритмами, проведем сравнение с наиболее популярными базовыми алгоритмами классификации: деревья решений, случайный лес, логистическая регрессия, наивный байесовский классификатор, нейронная сеть на усеченном наборе данных.

Оценка работоспособности алгоритмов была проведена при оптимальных настройках каждого алгоритма для одной и той же обучающей выборки

в 1000 точек. На рис. 3 показаны результаты сравнения качества классификации на указанном наборе данных в виде кривых ошибок (ROC-кривых) для оптимальных настроек каждого из базовых алгоритмов.



а) *б*) *в*) *г*) *д*) *е*)

Рис. 3. ROC-кривые качества классификации для деревьев решений (*а*), случайного леса (*б*), логистической регрессии (*в*), нейросетевой модели (*г*), байесовского классификатора (*д*) и джунглей решений (*е*)

В таблице 2 приведены основные метрики качества работы каждого из алгоритмов классификации: таблица сопряженности (матрица спутанности), которая показывает количество истинно положительных (True Positive – TP), ложноотрицательных (False Negative – FN), ложноположительных (False Positive – FP) и истинно отрицательных (True Negative – TN) результатов; доля правильных ответов алгоритма (Accuracy); доля объектов, названных классификатором положительными и при этом действительно являющихся положительными (Precision); Полнота – доля объектов положительного класса из всех положительно классифицированных объектов (Recall); F1-метрика – среднее гармоническое метрик Precision и Recall – позволяет оценить степень несбалансированности классов; площадь (Area Under Curve – AUC) под ROC кривой ошибок. Тестовая выборка состояла из 200 точек.

Таблица 2

Метрики качества классификации

Алгоритм	TP	FN	FP	TN	Accuracy	Precision	Recall	F1 Score	AUC
Деревья решений	1	10	6	183	0,920	0,143	0,091	0,111	0,671
Случайный лес	1	10	2	187	0,940	0,333	0,091	0,143	0,736
Логистическая регрессия	0	11	1	188	0,940	0,000	0,000	0,000	0,753
Нейросетевая модель	2	9	6	183	0,925	0,250	0,182	0,211	0,603
Байесовский классификатор	2	9	4	185	0,935	0,333	0,182	0,235	0,726
Графы (джунгли) решений	2	9	1	188	0,950	0,667	0,182	0,286	0,872

Заключение

Полученные результаты работы алгоритмов бинарной классификации при оптимальных настройках на усеченном наборе данных позволяют сделать следующие выводы:

1. Деревья решений на относительно малой выборке не позволяют создать достаточно точную модель, и, несмотря на то, что ансамблевые методы немного исправляют ситуацию, точностные характеристики оставляют желать лучшего.

2. Логистическая регрессия за счет умения работать с категориальными переменными показывает удовлетворительные и достаточно сбалансированные результаты, однако в отличие от остальных алгоритм классификации на основе логистической регрессии отработал с нулевым значением истинно положительных результатов, поэтому также имеет нулевые метрики Precision, Recall и F1 Score.

3. Нейросетевой алгоритм не успевает обучиться на относительно небольшом наборе данных и показывает худшие характеристики качества среди представленных алгоритмов.

4. Байесовский классификатор показывает наиболее близкие по сравнению с графами решений результаты, но при более низком значении метрики Precision, что говорит о худшей сбалансированности классов.

5. Джунгли решений на относительно небольшом наборе данных дают наилучшие результаты по всем метрикам точности: наибольшая площадь под ROC-кривой (AUC), что подтверждает их преимущества при работе как с пропущенными, так и с категориальными данными. Кроме того, это открывает перспективы для использования данного алгоритма для построения адаптивных алгоритмов классификации без использования ансамблевых методов.

Таким образом, в работе показаны основные преимущества и возможности применения графов принятия решений для задач бинарной классификации. Последовательное обучение DAG уровень за уровнем выполняется путем совместной оптимизации целевой функции как по выбору функции разделения, так и по структуре DAG. Были приведены сравнительные оценки качества работы классификационного алгоритма на примере задачи кредитного скоринга, при этом графы принятия решений показали себя наиболее эффективным алгоритмом принятия решений, который дал наилучшие результаты на небольшой обучающей выборке. Следует отметить, что графы (джунгли) решений могут найти эффективное применение для разнообразных и сложных задач классификации за счет повышения эффективности описательной и обобщающей способности по сравнению с классическим алгоритмом на основе деревьев решений, а также другими более сильными классификационными алгоритмами, которые показывают свою несостоятельность при недостатке данных для обучения.

Список литературы

1. Шитиков В. К., Мастицкий С. Э. Классификация, регрессия и другие алгоритмы Data Mining с использованием R. Тольятти ; Лондон, 2017. 351 с.
2. Mishra B. K., Hazra D., Tarannum K., Kumar M. Business Intelligence using Data Mining techniques and Business Analytics // 5th International Conference on System Modeling & Advancement in Research Trends (SMART 2016) (25–27 November 2016, Moradabad, India). Moradabad, 2016. P. 84–89. doi: 10.1109/SYSMART.2016.7894496
3. Мастицкий С. Э. Анализ временных рядов с помощью R. 2020. URL: https://ranalytics.github.io/tsa-with-r (дата обращения: 26.11.2020).
4. De’ath G. Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships // Ecology. 2002. № 83. P. 1105–1117.
5. Criminisi A., Shotton J. Decision Forests for Computer Vision and Medical Image Analysis // Proceedings of the 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and Sixth European Conference on Computational Biology. Springer, 2013. doi:10.1007/978-1-4471-4929-3_10.
6. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R. Springer, 2013. 436 p.
7. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: Data mining, inference, and prediction. Second Edition. Springer, 2009. 764 p.
8. Elisha O., Dekel S. Wavelet decompositions of random forests: smoothness analysis, sparse approximation and applications // J. Mach. Learn. Res. 2016. Vol. 17 (1). P. 6952–6989.
9. Decision jungles: compact and rich models for classification / J. Shotton, T. Sharp, P. Kohli, S. Nowozin [et al.] // Advances in Neural Information Processing Systems. 2013. P. 234–242.
10. Begon J. M., Joly A., Geurts P. Globally induced forest: a prepruning compression scheme // International Conference on Machine Learning. 2017. P. 420–428.

11. Decision Forests, Convolutional Networks and the Models in-Between / Y. A. Ioannou, D. Robertson, D. Zikic, P. Kontschieder [et al.]. 2016 ArXiv, abs/1603.01250. 2016.
12. Random Decision DAG: An Entropy Based Compression Approach for Random Forest / X. Liu, X. Liu, Y. Lai, F. Yang, Y. Zeng // Database Systems for Advanced Applications. DASFAA 2019. Lecture Notes in Computer Science / ed. by G. Li, J. Yang, J. Gama, J. Natwichai, Y. Tong. 2019. Vol. 11448. Springer, Cham. https://doi.org/10.1007/978-3-030-18590-9_37
13. Kislyakov A. N. Structuring advertising campaign costs considering the asymmetry of users' interests // Business Informatics. 2020. Vol. 14, № 4. P. 7–18. doi:10.17323/2587-814X.2020.4.7.18.
14. Kislyakov A., Tikhonuyk N. Principles for Development of Predictive Stability Models of Social and Economic Systems on the basis of DTW // First Conference on Sustainable Development: Industrial Future of Territories (IFT 2020). 2020. Vol. 208, № 08001. doi:10.1051/e3sconf/202020808001.
15. Королев О. Л., Куссый М. Ю., Сигал А. В. Применение энтропии при моделировании процессов принятия решений в экономике / под ред. А. В. Сигала. Симферополь : ОДЖАКЪ, 2013. 148 с.

References

1. Shitikov V.K., Mastitskiy S.E. *Klassifikatsiya, regressiya i drugie algoritmy Data Mining s ispol'zovaniem R = Classification, regression, and other Data Mining algorithms using R.* Tolyatti; London, 2017:351. (In Russ.)
2. Mishra B.K., Hazra D., Tarannum K., Kumar M. Business Intelligence using Data Mining techniques and Business Analytics. *5th International Conference on System Modeling & Advancement in Research Trends (SMART 2016) (25–27 November 2016, Moradabad, India)*. Moradabad, 2016:84–89. doi:10.1109/SYSMART.2016.7894496.
3. Mastitskiy S.E. *Analiz vremennykh ryadov s pomoshch'yu R = Time series analysis using R*. 2020. (In Russ.). Available at: https://ranalytics.github.io/tsa-with-r (accessed 26.11.2020).
4. De'ath G. Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships. *Ecology*. 2002;83:1105–1117.
5. Criminisi A., Shotton J. Decision Forests for Computer Vision and Medical Image Analysis. *Proceedings of the 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and Sixth European Conference on Computational Biology*. Springer, 2013. doi:10.1007/978-1-4471-4929-3_10.
6. James G., Witten D., Hastie T., Tibshirani R. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013:436.
7. Hastie T., Tibshirani R., Friedman J. *The elements of statistical learning: Data mining, inference, and prediction*. Second Edition. Springer, 2009:764.
8. Elisha O., Dekel S. Wavelet decompositions of random forests: smoothness analysis, sparse approximation and applications. *J. Mach. Learn. Res*. 2016;17(1):6952–6989.
9. Shotton J., Sharp T., Kohli P., Nowozin S. [et al.]. Decision jungles: compact and rich models for classification. *Advances in Neural Information Processing Systems*. 2013:234–242.
10. Begon J.M., Joly A., Geurts P. Globally induced forest: a prepruning compression scheme. *International Conference on Machine Learning*. 2017:420–428.
11. Ioannou Y. A., Robertson D., Zikic D., Kontschieder P. [et al.]. *Decision Forests, Convolutional Networks and the Models in-Between*. 2016. ArXiv, abs/1603.01250. 2016.
12. Liu X., Liu X., Lai Y., Yang F., Zeng Y. Random Decision DAG: An Entropy Based Compression Approach for Random Forest. *Database Systems for Advanced Applica-*

tions. DASFAA 2019. Lecture Notes in Computer Science. 2019;11448. Springer, Cham. https://doi.org/10.1007/978-3-030-18590-9_37
13. Kislyakov A.N. Structuring advertising campaign costs considering the asymmetry of users' interests. *Business Informatics*. 2020;14(4):7–18. doi:10.17323/2587-814X.2020.4.7.18.
14. Kislyakov A., Tikhonuyk N. Principles for Development of Predictive Stability Models of Social and Economic Systems on the basis of DTW. *First Conference on Sustainable Development: Industrial Future of Territories (IFT 2020)*. 2020;208(08001). doi:10.1051/e3sconf/202020808001.
15. Korolev O.L., Kussyy M.Yu., Sigal A.V. *Primenenie entropii pri modelirovanii protsessov prinyatiya resheniy v ekonomike = Application of entropy in modeling decision-making processes in the economy*. Simferopol: ODZhAK", 2013:148. (In Russ.)

Информация об авторах / Information about the authors

Алексей Николаевич Кисляков
кандидат технических наук, доцент
кафедры информационных технологий,
Владимирский филиал Российской
академии народного хозяйства
и государственной службы
при Президенте Российской Федерации
(Россия, г. Владимир, ул. Горького, 59а)
E-mail: ankislyakov@mail.ru

Alexey N. Kislyakov
Candidate of technical sciences,
associate professor of sub-department
of information technology,
Vladimir branch of the Russian Academy
of National Economy and Public
Administration under the President
of the Russian Federation
(59a, Gorky street, Vladimir, Russia)