С.М. Гордеева, В.Н. Малинин

ИСПОЛЬЗОВАНИЕ DATA MINING В ЗАДАЧЕ ГИДРОМЕТЕОРОЛОГИЧЕСКОГО ПРОГНОЗИРОВАНИЯ

S.M. Gordeeva, V.N. Malinin

USE OF DATA MINING IN HYDROMETEOROLOGICAL FORECASTING

Представлено краткое описание нового метода анализа экспериментальных данных Data Mining и его направления — деревья решений. Показана эффективность использования алгоритма CART для долгосрочного прогноза годового стока р. Северной Двины. Стандартная ошибка прогностических значений стока по независимым данным составляет 0,65 от величины стандартного отклонения фактических значений годового стока.

Ключевые слова: Data Mining, деревья решений, алгоритм CART, прогноз годового стока, Северная Двина.

A brief description of Data Mining and its branch — decision trees as a new method of experimental data analysis is presented. Application of the CART algorithm for the long-term forecast of the annual runoff of the Severnaya Dvina river showed its efficiency. The standard error of the forecast for independent data series on runoff is 0.65 of the value of the annual runoff standard deviation.

Keywords: Data Mining, decision trees, CART algorithm, annual runoff forecast, Severnaya Dvina river.

Дословно Data Mining (discovery-driven data mining) переводится как «добыча» (раскопка) данных. По сути Data Mining — анализ экспериментальных данных с целью поиска неочевидных, объективных и полезных на практике закономерностей. Неочевидных — это значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем. Объективных — что обнаруженные закономерности будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным. Полезных — это значит, что выводы имеют конкретное значение, которому можно найти практическое применение [9]. На русском языке этот вид анализа еще не получил устоявшейся терминологии. Data Mining является мультидисциплинарной областью, возникшей и развивающейся на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, нейронных сетей, теории баз данных и др. Отсюда обилие методов и алгоритмов, реализованных в различных действующих системах Data Mining [1, 4, 5, 7, 8, 11, 13, 14].

К методам и алгоритмам Data Mining относятся: искусственные нейронные сети, деревья решений, символьные правила, метод опорных векторов, байесовские сети, линейная регрессия, корреляционно-регрессионный анализ, методы кластерного анализа, методы поиска ассоциативных правил, метод ограниченного перебора,

эволюционное программирование и генетические алгоритмы, разнообразные методы визуализации данных и множество других методов. В связи с этим удобно все многообразие методов Data Mining разделять на две группы: *статистические* и *кибернетические*.

Довольно долго Data Mining не признавалась полноценной самостоятельной областью анализа данных, иногда ее даже называли «задворками статистики» [15]. Однако в настоящее время идет ее интенсивное развитие. В основу современной технологии Data Mining положена концепция шаблонов (паттернов), отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные подвыборкам данных, которые могут быть компактно выражены в понятной человеку форме. Поиск шаблонов производится методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей. Важное положение Data Mining — нетривиальность разыскиваемых шаблонов. Это означает, что найденные шаблоны должны отражать неочевидные, неожиданные (unexpected) регулярности в данных, составляющие так называемые скрытые знания (hidden knowledge). К обществу пришло понимание, что сырые данные (raw data) содержат глубинный пласт знаний, при грамотной раскопке которого могут быть обнаружены настоящие самородки [3].

К числу наиболее популярных методов Data Mining принадлежит *метод деревьев решений* (decision trees), который может быть использован в практике многих гидрометеорологических задач, в частности задач классификации и прогнозирования. Если зависимая переменная принимает дискретные значения, то при помощи метода дерева решений решается задача классификации. Если зависимая переменная принимает непрерывные значения, то дерево решений устанавливает зависимость этой переменной от независимых переменных, то есть решает задачу численного прогнозирования. Очевидно, впервые суть решений деревьев была изложена в 1966 г. в работе Ханта и др. [10].

К достоинствам деревьев решений относятся:

- визуализация получаемых результатов и более понятная их интерпретация;
- алгоритм конструирования дерева решений не требует от пользователя выбора независимых переменных. На вход алгоритма можно подавать все переменные, и алгоритм сам выберет наиболее значимые из них, которые в дальнейшем будут использованы для построения дерева;
- большинство алгоритмов конструирования деревьев решений имеют возможность специальной обработки пропущенных значений;
- точность моделей, созданных при помощи решений деревьев, сопоставима с другими методами построения классификационных моделей (статистические методы, нейронные сети);
- деревья решений могут работать с любыми типами исходных данных, как с числовыми, так и с категориальными переменными;
- деревья решений строят непараметрические модели, то есть исходные данные свободны от теоретического распределения.

В наиболее простом виде дерево решений — это способ представления классифицирующих правил в виде иерархической структуры. Ее основой, имеющей вид дерева, являются правила типа «*если... то...*» (*if – then*). Для принятия решения, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие

в узлах этого дерева, начиная с его корня. Корень — исходный вопрос, внутренний узел дерева, является узлом проверки определенного условия. Далее идет следующий вопрос и т.д., пока не будет достигнут конечный узел дерева, являющийся узлом решения. Вопросы имеют вид «*значение параметра А больше х?*». Если ответ «да», то осуществляется переход к правому узлу следующего уровня, если «нет» — то к левому узлу; затем снова следует вопрос, связанный с соответствующим узлом.

В соответствии с иерархической природой деревьев ветвления выполняются последовательно, начиная с корневой вершины, затем переходят к вершинам потомков до тех пор, пока дальнейшее ветвление не прекратится и «неразветвленные» вершины потомки окажутся терминальными. Терминальные вершины («листья») — это узлы дерева, начиная с которых никакие решения больше не принимаются. В программе Statistica, например на рисунке дерева, решения терминальные вершины выделяются красными пунктирными линиями, а остальные — так называемые решающие вершины, или вершины ветвления, — сплошными синими линиями. Началом дерева считается самая верхняя решающая вершина, которую иногда также называют корнем дерева.

В программе Statistica [12] предусмотрено три алгоритма ветвления дерева.

Первый алгоритм — *дискриминантное одномерное ветвление* для категориальных и порядковых предикторов (QUEST) — можно использовать, если независимые переменные являются категориальными, порядковыми или смесью обоих типов. При этом под порядковыми переменными понимаются любые количественные переменные.

Второй алгоритм — *дискриминантное ветвление по линейным комбинациям порядковых предикторов* — можно применять, если для анализа выбраны только порядковые переменные.

Третий алгоритм — *полный перебор для одномерных ветвлений* методом C&RT (CART) — можно использовать для категориальных, порядковых или смеси обоих типов независимых переменных. По сравнению с предыдущими методами ветвления в этом методе для нахождения наилучшего варианта выполняется последовательный перебор всех возможных комбинаций независимых переменных. В связи с тем, что количество вариантов для такого перебора может оказаться очень большим, вычисления могут занять много времени, вследствие чего дерево решений окажется достаточно сложным.

Процесс создания дерева происходит сверху вниз, то есть является нисходящим. Алгоритмы конструирования деревьев решений состоят из этапов «построение», или «создание», дерева (tree building) и «сокращение» дерева (tree pruning). В ходе создания дерева решаются вопросы выбора критериев расщепления (разбиения), ветвления и остановки обучения. При сокращении дерева решается вопрос определения подходящего размера дерева, то есть отсечения некоторых его ветвей. Впрочем, деление на отдельные этапы во многом условно, поскольку некоторые алгоритмы выполняют построение дерева и его сокращение последовательно, а другие — чередуют их в процессе своей работы для предотвращения наращивания внутренних узлов.

Критерий расщепления требуется для того, чтобы разбить множество на однородные подмножества, которые должны быть связаны с данным узлом. Каждый узел помечается определенным атрибутом. Существует правило выбора атрибута: он должен разбивать исходное множество данных таким образом, чтобы объекты подмножеств, получаемых в результате этого разбиения, являлись представителями одного класса

или же были максимально приближены к такому разбиению. Это означает, что в каждом классе количество инородных объектов из других классов («примесей») должно быть минимальным.

Существуют различные критерии расщепления [3]. Наиболее известные — мера энтропии и индекс *Gini*. Мера энтропии в некоторых алгоритмах известна под названием «мера информационного выигрыша» (information gain measure). Другой критерий расщепления реализован в алгоритме CART (Classification and Regression Tree). Как видно из названия, он решает задачи классификации и регрессии. Алгоритм был разработан в 1974–1984 гг. четырьмя профессорами статистики ведущих американских университетов: Л. Брейманом, Д. Фридманом, Ч. Стоуном и Р. Олшеном [6]. В алгоритме CART в качестве критерия расщепления используется индекс *Gini*. При помощи этого индекса атрибут выбирается на основании расстояний между распределениями классов. Индекс *Gini* определяется по формуле

$$Gini(T) = 1 - \sum_{j=1}^{n} p_j^2, \qquad (1)$$

где T — текущий узел; p_j — вероятность класса j в узле T; n — количество классов.

Данная оценочная функция основана на идее уменьшения неопределенности в узле. Допустим, есть узел, и он разбит на два класса. Максимальная неопределенность в узле будет достигнута при разбиении его на два подмножества по 50 примеров, а максимальная определенность — при разбиении на 100 и 0 примеров.

Алгоритм CART предназначен для построения бинарного дерева решений. Бинарные деревья также называют двоичными. Каждый узел бинарного дерева при разбиении имеет только двух потомков, называемых дочерними ветвями. Дальнейшее разделение ветви зависит от того, много ли исходных данных описывает данная ветвь. На каждом шаге построения дерева правило, формируемое в узле, делит заданное множество примеров на две части. Правая его часть (ветвь right) — это та часть множества, в которой правило выполняется; левая (ветвь left) — та, для которой правило не выполняется.

Важно, что алгоритм CART работает как с числовыми, так и с категориальными атрибутами. В каждом узле разбиение может идти только по одному атрибуту. Если атрибут является числовым, то во внутреннем узле формируется правило вида $x_i <= c$. Значение c в большинстве случаев выбирается как среднее арифметическое двух соседних упорядоченных значений переменной x_i обучающего набора данных. Если же атрибут относится к категориальному типу, то во внутреннем узле формируется правило $x_i < V(x_i)$, где $V(x_i)$ — некоторое непустое подмножество множества значений переменной x_i в обучающем наборе данных.

С точки зрения качества получаемых результатов, весьма важной представляется оценка точности распознавания. В методе decision trees наиболее точной классификацией считается та, которая связана с наименьшей ценой. При этом *под ценой понимается доля неправильно классифицированных объектов*. Под неправильной классификацией понимается ошибочное отнесение объекта, принадлежащего одному классу, в другой класс. Очевидно, что самым лучшим прогнозом будет тот, который дает наименьший процент неправильных классификаций.

В модуле *Деревья классификации* программы Statistica [12] предусмотрено два варианта задания цены неправильной классификации объектов: *Равные* и *Пользовательские*. В первом варианте, когда цены берутся одинаковые, все внедиагональные элементы матрицы цен ошибок классификации (прогнозируемые классы — по строкам, наблюдаемые классы — по столбцам) полагаются равными 1. Второй вариант соответствует случаю, когда по каким-либо причинам для одних классов требуется более точный прогноз, чем для других. Наиболее просто это можно учесть путем придания «важным» классам больших «весов», чем остальным.

В рамке *Априорные вероятности* нужно указать способ вычисления априорных вероятностей, то есть задать вероятность того, что объект будет принадлежать определенному классу даже при отсутствии какой-либо априорной информации о значениях независимых переменных. Если при этом априорные вероятности выбраны пропорциональными численности классов, а цена ошибки классификации является одинаковой для всех классов, то минимизация потерь будет в точности эквивалентна минимизации доли неправильно классифицированных наблюдений. Отметим, что выбор априорных вероятностей, используемых для минимизации потерь, очень сильно влияет на результаты классификации. Если различия между исходными частотами в данной задаче не считаются существенными или если мы заранее знаем, что классы содержат примерно одинаковое количество наблюдений, то тогда можно взять одинаковые априорные вероятности, что соответствует опции *Равные*. Если исходные частоты связаны с размерами классов, то следует в качестве оценок для априорных вероятностей взять относительные размеры классов в выборке опция *Оцениваемые*. Наконец, если мы располагаем какой-то информацией об исходных частотах, то априорные вероятности нужно выбирать с учетом этой информации в опции *Пользовательские*.

Весьма сложным и неоднозначным является вопрос остановки ветвления дерева, который одновременно можно рассматривать также с точки зрения определения подходящего размера дерева. Вообще говоря, понятно, что, чем больше размерность дерева классификации, тем точнее прогноз, но одновременно сложнее интерпретация результатов и решающие правила, поэтому пользователю труднее сделать прогноз о принадлежности к классу нового объекта. В связи с этим можно лишь высказать ряд общих соображений о том, что следует считать «подходящими размерами» для дерева классификации. С одной стороны, дерево классификации должно быть достаточно сложным, чтобы учитывать имеющуюся информацию, а с другой — оно должно быть как можно проще, с точки зрения интерпретации получаемых результатов. Такие взаимоисключающие требования к дереву приводят к необходимости введения каких-либо критериев оптимальности, позволяющих максимальным образом учитывать эти требования.

Отметим также, что дерево должно уметь использовать ту информацию, которая улучшает точность прогноза, и игнорировать информацию, которая прогноз не улучшает. Одна из возможных стратегий выбора размера дерева состоит в том, чтобы наращивать его до нужного размера, который определяется самим пользователем на основе имеющихся данных, выданных на предыдущих этапах анализа или, в крайнем случае, на основе интуиции. В принципе, если не установлено ограничение на число ветвлений, то можно прийти к «чистой» классификации, когда каждая терминальная

вершина содержит только один класс наблюдений (объектов). Следует также иметь в виду, что привлекаемые к классификации исходные данные содержат ошибки измерений, причем степень «засоренности» ими выборки, как правило, не известна. Поэтому нецелесообразно осуществлять сортировку до тех пор, пока каждая терминальная вершина станет «чистой».

Алгоритм CART принципиально отличается от других алгоритмов конструирования деревьев решений механизмом, имеющим название minimal cost-complexity tree pruning. В рассматриваемом алгоритме отсечение — это некий компромисс между получением дерева «подходящего размера» и получением наиболее точной оценки классификации. Метод заключается в получении последовательности уменьшающихся деревьев, но деревья рассматриваются не все, а только их «лучшие представители».

В модуле CART программы Statistica [12] на вкладке *Параметры остановки* реализованы два варианта остановки: *По отклонению* и *Прямая остановка (FACT)*. По первому правилу отсечения ветви дерева последовательно «отсекаются» от полного дерева классификации. Затем дерево классификации «подходящего размера» выбирается из «усеченных» деревьев с помощью специального правила стандартной ошибки. При выборе этого правила в рамке *Условия остановки* активизированы опции *Максимальное n* (число узлов) дерева и *Минимальное число случаев* на вершине (листе), при помощи которых пользователь может управлять моментом остановки ветвлений дерева.

В правиле *Прямая остановка по методу FACT* используется совершенно иной подход. Здесь полное дерево классификации, содержащее все возможные ветвления, рассматривается как имеющее «подходящий размер». При этом правиле надо определить момент прекращения ветвлений, задавая в рамке *Условия остановки* значение параметра, имеющего название *Доля неклассифицированных объектов*. В этом случае ветвление по независимым переменным будет продолжаться до тех пор, пока каждая терминальная вершина дерева классификации не станет «чистой» или количество отнесенных к этой вершине объектов из одного или нескольких классов не станет меньше заданной доли от общей численности соответствующего класса (классов).

Стратегия выбора «подходящего размера» для дерева — метод автоматического построения дерева по Брейману [6], который реализован в виде *кросс-проверочного отсечения* по *минимальной цене-сложности*, либо по *минимальному отклонению-сложности*. Единственное различие между этими способами — метод измерения ошибки прогноза. При первой опции используется функция потерь, равная доли неправильно классифицированных объектов при оцениваемых априорных вероятностях и одинаковых ценах ошибок классификации. При второй опции используется мера, основанная на принципе максимума правдоподобия. Для того чтобы в модуле деревья классификации выполнить кросс-проверочное отсечение по минимальной цене-сложности, нужно выбрать правило остановки по ошибке классификации. Кросс-проверочное отсечение по минимальному отклонению-сложности выполняется, если выбрано правило остановки по отклонению.

Кросс-проверка определяется путем расчета доли неправильно классифицированных объектов на независимой выборке (цена кросс-проверки, *CV-cost*). В модуле CART программы Statistica [12] — настраивается на вкладке *Проверка*, где предусмотрено два способа выбора независимой выборки:

1. *Кросс-проверка на тестовой выборке* — наиболее предпочтительный вариант кросс-проверки. В этом варианте дерево классификации строится по исходной обучающей выборке, а его способность к прогнозированию проверяется путем предсказания классовой принадлежности элементов по независимой (тестовой) выборке. Если значение цены по независимой выборке окажется больше, чем по обучающей выборке, то это свидетельствует о плохом результате кросс-проверки. Возможно, в этом случае следует поискать дерево другого размера, которое бы лучше выдерживало кросс-проверку. Независимая и обучающая выборки могут быть сформированы, например, путем деления исходной выборки на две не обязательно равные части.
2. *V-кратная кросс-проверка*. Этот вид кросс-проверки целесообразно использовать в случаях, когда в распоряжении пользователя нет отдельной тестовой (независимой) выборки, а обучающее множество слишком мало для того, чтобы из него выделять тестовую выборку. В этом варианте производится заданное число итераций (V), причем всякий раз часть обучающей выборки (равная единице, деленной на заданное целое число) оставляется в стороне, а затем по очереди каждая из отложенных частей используется как тестовая выборка для кросс-проверки построенного дерева классификации. По умолчанию число итераций равно 10.

В настоящее время разработано уже довольно много алгоритмов decision trees. Для решения многих гидрометеорологических задач, в том числе классификации и прогнозирования, на наш взгляд, наиболее перспективным является алгоритм CART. Основные характеристики алгоритма CART: бинарное расщепление, критерий расщепления — индекс *Gini*, алгоритмы minimal cost-complexity tree pruning и V-fold cross-validation, принцип «вырастить дерево, а затем сократить», высокая скорость построения, обработка пропущенных значений и робастность, то есть устойчивость к шумам и выбросам данных.

Рассмотрим тестовый пример функцией отклика — категориальной переменной. По данным за 15 ч в Санкт-Петербурге за июнь — август 2013 г. ($N = 92$) была составлена матрица следующих атмосферных параметров: атмосферное давление, температура воздуха, влажность, направление и скорость ветра, общее количество облаков, облачность нижнего, среднего и верхнего ярусов, то есть всего 9 характеристик. С помощью построения дерева решалась классификационная задача: при каких атмосферных условиях идет дождь, а при каких — нет. Из исходной матрицы следует, что из 92-х дней дождь шел 19 раз, что соответствует 21 % случаев и 73 дня (79 % случаев) он отсутствовал.

Построение дерева осуществлялось на основе алгоритма CART. Вначале было построено полное дерево, которое, как и следовало ожидать, имело довольно сложный вид. В связи с этим возникла задача выбора оптимального дерева. В алгоритме CART в соответствии с [12] его автоматический выбор осуществляется следующим образом: берется дерево с минимальной величиной ценой кросс-проверки (*CV-cost*) и к нему прибавляется его стандартная ошибка (*CV std. error*). Эта сумма принимается за пороговую величину $CV_{крит}$. Далее осуществляется перебор в сторону деревьев меньшей сложности и оптимальным считается такое последнее дерево, для которого

его величина *CV-cost* меньше $CV_{крит}$. Таким образом, выполняется требование оптимальности параметра «сложность/цена».

В табл. 1 приводятся оценки кросс-проверки для четырех деревьев. Отметим, что деревья в программе Statistica [12] нумеруются от более сложных к простым. Полное дерево, обозначенное цифрой 1, включает в себя те исходные переменные, которые участвуют в процессе ветвления с целью выявления условий, когда идет дождь, а когда — нет. Дерево 1 имеет 6 терминальных вершин, непосредственно участвующих в ветвлении, и 5 нетерминальных вершин, которые в данном дереве в ветвлении не участвуют. Итого общее число вершин равно 11. Следующее дерево 2 имеет 3 терминальные вершины и 2 нетерминальные. Наконец, последнее дерево 4 имеет только 1 терминальную вершину, то есть представляет собой исходную выборку.

Таблица 1

Оценки кросс-проверки деревьев различной сложности для комплекса атмосферных параметров в Санкт-Петербурге за летний период (июнь – август) 2013 г.

Показатель	Число терминальных вершин (*Terminal Nodes*)	Цена кросс-проверки (*CV cost*)	Стандартная ошибка кросс-проверки (*CV std. error*)	Цена проверки на обучающей выборке (*Resubstitution cost*)	Сложность узла (*Node complexity*)
Дерево 1	6	0,228	0,043	0,032	0,000
Дерево 2	3	0,206	0,042	0,086	0,018
Дерево 3	2	0,184	0,040	0,108	0,021
Дерево 4	1	0,206	0,042	0,206	0,097

Как видно из табл. 1, минимальная оценка *CV cost* отмечается для дерева 3. С учетом стандартной ошибки $CV_{крит} = 0{,}224$. Далее мы должны двигаться в сторону более простых деревьев и искать такое дерево, для которого выполняется условие $CV\text{-}cost < CV_{крит}$. Этому условию в табл. 1 соответствует дерево 4, то есть непосредственно исходная выборка. Однако по представленным в ней параметрам невозможно определить метеорологические условия, при которых идет дождь, а при которых — нет.

Итак, критерий оптимальности в программе Statistica в данном случае оказывается неинформативным, поэтому мы решили использовать другую возможную стратегию выбора размера дерева, состоящую в том, чтобы наращивать его до такого размера, который бы наилучшим образом соответствовал имеющимся в исходной выборке данным.

На рис. 1 приводятся результаты первого ветвления для атмосферных параметров в Санкт-Петербурге за летний период (июнь – август) 2013 г. Нетрудно видеть, что первым параметром, разделяющим условия «идет дождь» и «нет дождя», является не атмосферное давление, как можно было предположить, а влажность воздуха. Если относительная влажность (U) меньше 68,5 %, то дождь отсутствует, если больше — дождь идет. При $U < 68{,}5$ % имеем 75 случаев, из них 69 дней дождя не было, а в течение 6 дней дождь шел. Условие $U > 68{,}5$ % соответствует 17 дням, из которых 13 дней шел дождь, а 4 дня он отсутствовал.

Итак, влажность воздуха дает правильное решение выполнения условия «идет дождь» и «нет дождя» в 89 % случаях (рис. 2), то есть цена неправильной классификации

составляет 11 %. Однако если рассматривать эти условия по отдельности, то ошибки существенно различаются. Так, ошибка «нет дождя», когда он идет, составляет 8 %, в то время как ошибка условия «идет дождь», когда его нет, равна 23,5 %.

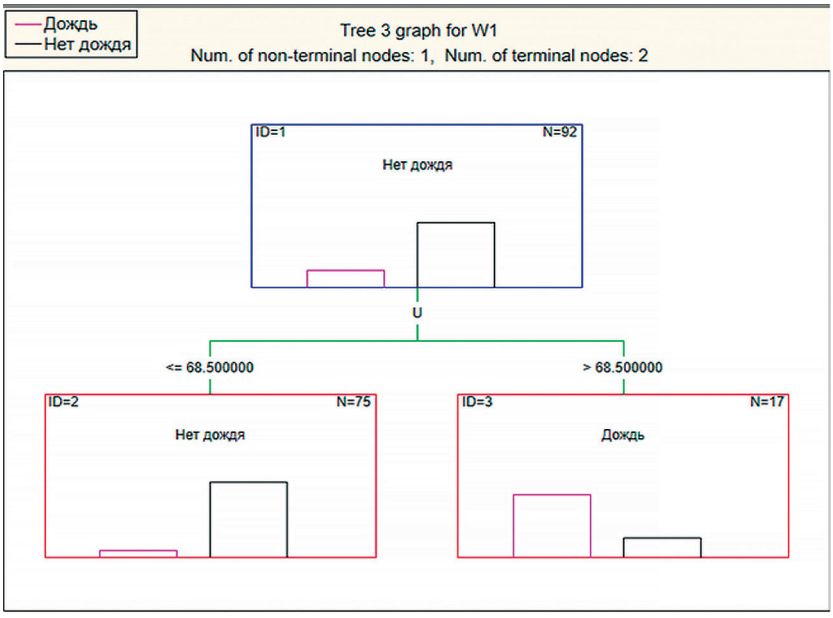

Рис. 1. Дерево решений после первого ветвления для атмосферных параметров в Санкт-Петербурге за летний период (июнь – август) 2013 г. с целью выявления условий, при которых дождь идет или не идет

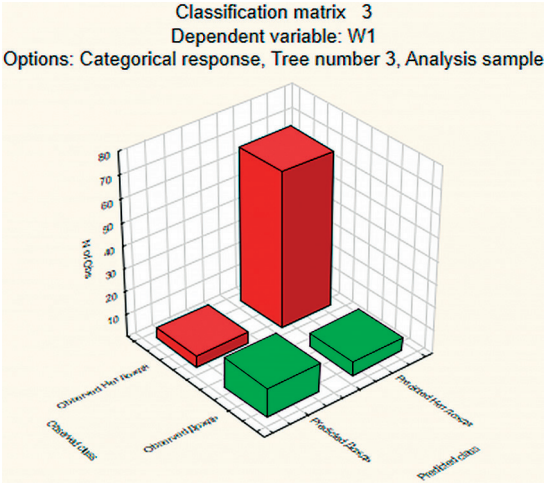

Рис. 2. Результаты разбиения алгоритмом CART атмосферных параметров в Санкт-Петербурге за летний (июнь — август) период 2013 г. при пороговом значении влажности воздуха 68,5 %

Естественно полагать, что при дальнейшем ветвлении качество разбиения может быть улучшено. При этом ввиду малых ошибок левая ветвь дерева вряд ли нуждается в уточнении. Очевидно, имеет смысл рассматривать ветвление только правой части дерева. Разделителем выборки из 17 дней с максимальной вероятностью дождя ($ID = 3$) является атмосферное давление (P). При $P < 756{,}8$ мм рт. ст. выявляется 11 случаев с дождем, причем дней без дождей не было. При $P > 756{,}8$ мм рт. ст. имеем 6 случаев, из которых в действительности дождь шел только 2 раза, а в течение 4 дней он отсутствовал. Итак, имеем полное совпадение с вариантом $U > 68{,}5$ %, когда дождь шел 13 дней, а 4 дня его не было. Следовательно, ветвление дерева с атмосферным давлением уже является лишним. Отсюда следует, что в качестве оптимального дерева можно принять результаты после первого ветвления, то есть дерево 3, содержащее 2 терминальные вершины, в котором единственным разделителем служит влажность воздуха.

Рассмотрим задачу долгосрочного прогноза годового стока р. Сев. Двина — числовой переменной. Отметим, что прогноз стока этой реки очень сложен из-за слабого покрытия речного водосбора гидрометеорологическими данными. Как известно, определяющими факторами межгодовых колебаний стока крупных рек являются: а) запасы влаги в снежном покрове перед началом снеготаяния; б) предшествующее осеннее увлажнение почвы; в) предшествующее летнее увлажнение в апреле — сентябре. При этом, чем больше площадь водосборного бассейна, тем больший вклад этих факторов в колебания стока и тем более длительную предысторию стокоформирующих факторов следует учитывать. Это означает, что сток в i-й год зависит не только от увлажнения в $i-1$ год, но и частично в $i-2$ год.

Итак, основная рабочая гипотеза может быть сформулирована следующим образом. Накопление влаги (общее увлажнение) в бассейне за два предшествующих началу половодья года практически полностью определяет речной сток в его замыкающем створе до начала следующего половодья. Естественно, главное влияние на сток оказывает первый предшествующий год. Влияние второго года сказывается главным образом в аномальные по характеру увлажнения годы. Учтем также, что межгодовая изменчивость осадков значительно превышает аналогичную изменчивость суммарного испарения. Это означает, что испарением можно пренебречь без существенной потери точности в расчетах. В результате прогностическая модель стока в i-й год может быть записана в следующем виде:

$$Q_i = f\left\{\sum^{k1} P_{i-1}, \sum^{k2} (P)_{i-1}, \sum^{k1} P_{i-2}, \sum^{k2} (P)_{i-2}\right\}, \qquad (2)$$

где k_1, k_2 — количество месяцев в холодный (октябрь–март) и теплый (апрель–сентябрь) периоды года соответственно. Но поскольку в гидрологических архивах сток рек обычно дается для календарных годовых периодов, то перейдем в формуле (2) к годовому стоку, а в качестве последнего $(i-1)$ периода увлажнения примем теплое полугодие. В этом случае минимальная заблаговременность прогноза стока в формуле (2) будет составлять 15 месяцев.

Годовой сток Северной Двины брался за период 1982–2012 гг. Осадки за указанный период времени доступны только для 8 станций водосбора реки (http://aisori.meteo.

ru/ClimateR). В соответствии с формулой (2) был сформирован архив из 32 временных рядов предикторов. При этом архив был разделен на 2 части: зависимую выборку (1982–2008), которая использовалась для построения прогностической модели, и независимую выборку (2009–2012), по которой выполнялось сравнение фактических и вычисленных по модели значений годового стока.

Вначале для прогноза стока применялся физико-статистический метод, основой которого является модель множественной линейной регрессии (МЛР). С помощью пошагового алгоритма МЛР [2] получена оптимальная модель со следующими статистическими параметрами: число переменных — 6, коэффициент детерминации — 0,74, критерий Фишера — 8,9, стандартная ошибка — 236 м³/с. В принципе, для зависимой выборки результаты неплохие. Однако по независимой выборке стандартная ошибка прогноза составила 427 м³/с, что даже превышает стандартное отклонение величины годового стока (402 м³/с). Это означает, что при таком слабом покрытии территории бассейна осадкомерными станциями данная регрессионная модель не позволяет прогнозировать годовой сток Северной Двины, и необходим поиск альтернативных способов прогноза.

Поэтому мы обратились к методу деревьев решений. Моделирование стока реки Северная Двина выполнялось алгоритмом CART с априорными вероятностями, пропорциональными численности классов и ценой ошибки классификации одинаковой для всех классов. Ограничение на число ветвлений устанавливалось по отклонению с V-кратной кросс-проверкой ($V = 10$), что обеспечивает отсечение дерева по минимальной цене-сложности. Исходными данными послужила выборка из 32 временных рядов предикторов.

На рис. 3 представлено распределение значений цены проверки на обучающей (зависимой) выборке (*Resubstitution cost*) и цены ошибки кросс-проверки (*CV cost*) в зависимости от номера дерева. Дерево 1 имеет 8 терминальных вершин, а последнее дерево 8 — только 1 вершину. Из рис. 3 следует, что с увеличением числа вершин ошибки обучения быстро уменьшаются. Так, уже после второго ветвления (дерево 6, 3 вершины) она уменьшилась почти в 3 раза, а после пятого ветвления (дерево 3, 6 вершин) — в 5 раз. Более сложным оказывается распределение ошибок кросс-проверки. Однако принципиально важным является тот факт, что минимальное значение *CV-cost* соответствует дереву 8 с одной вершиной. Это означает, что данное дерево решений следует считать оптимальным, однако оно практически не позволяет осуществлять прогноз стока. Как и в предыдущем случае, критерий оптимальности не срабатывает.

Для определения дерева оптимального размера мы использовали ту же стратегию, что и в предыдущем тестовом примере. Вначале рассматривали результаты прогноза стока Двины после первого ветвления, затем после второго и так далее, вплоть до последнего дерева. На первом ветвлении разделителем выступают летние осадки за предшествующий $i-1$ год в пункте Онега: если осадков выпало мало (< 341,2 мм), то прогнозируемый, средний для этих случаев, сток реки (3495 м³/с) выше нормы (3317 м³/с); если осадков выпало больше 341,2 мм, то ожидается годовой сток меньше нормы (2981 м³/с) (рис. 4). В первом случае имеем 17 значений стока, во втором — 9. На рис. 4 «Mu» означает среднее значение, а «Var» — дисперсию прогнозируемого ряда годового стока.

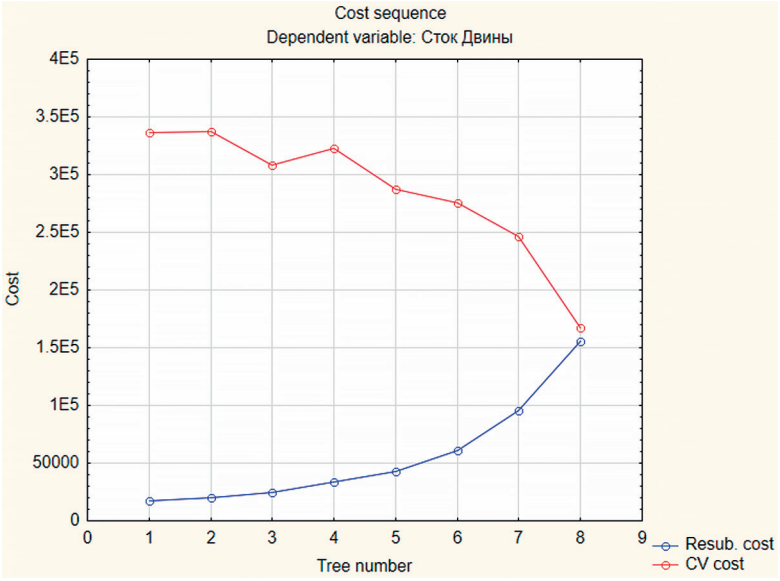

Рис. 3. Распределение значений цены проверки на обучающей (зависимой) выборке (*Resubstitution cost*) и цены ошибки кросс-проверки (*CV cost*) в зависимости от количества узлов дерева при выявлении связи годового стока Северной Двины с осадками на водосборе за два предшествующих года

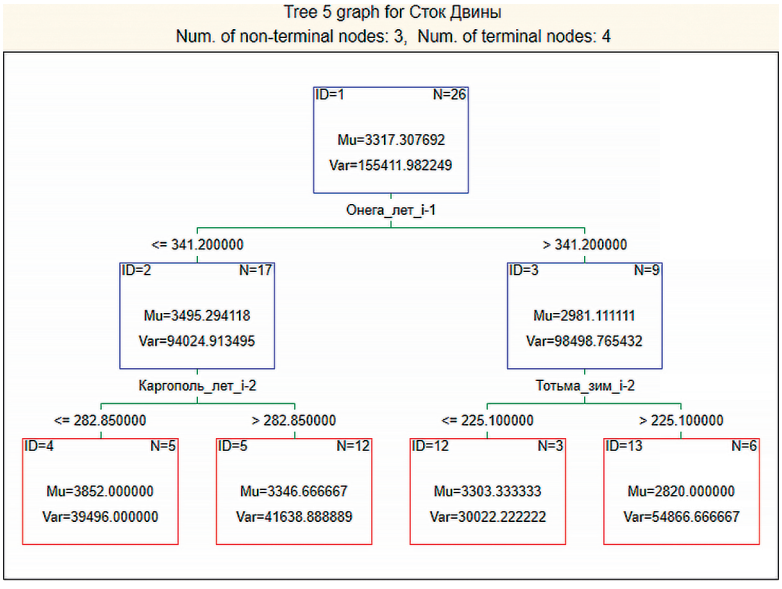

Рис. 4. Дерево решений 5, описывающее формирование годового стока Северной Двины в *i*-й год в зависимости от зимних и летних осадков в $i-1$ и $i-2$ годы на 8 метеорологических станциях, расположенных на территории бассейна

На втором ветвлении происходит уточнение формирования 17 значений повышенного стока Сев. Двины за счет летних осадков в пункте Каргополь за $i-2$ год: если осадков выпадает много (> 282,8 мм), то прогнозируется 12 значений годового стока (3347 м³/с) близкого к норме, если осадков было меньше 282,8 мм, то ожидается 5 значений стока (3852 м³/с), значительно превышающего норму. Выявление фактора формирования оставшихся 9 значений стока реки происходит лишь на 6 ветвлении. Разделителем выступают зимние осадки в пункте Тотьма за $i-2$ год. Если осадков было мало (< 225,1 мм) — прогнозируется умеренное состояние стока со средним значением 3303,3 м³/с, если больше — значительно меньший сток (среднее значение 2820,0 м³/с). Для последующих деревьев разделителями выступают зимние осадки в Шенкурске и в Койнасе за $i-2$ год, в Койнасе за $i-1$ год и летние осадки в Койнасе за $i-1$ год. Другие переменные из исходной выборки не дают какого-либо вклада в описание дисперсии годового стока Северной Двины.

После этого были выполнены расчеты коэффициента детерминации и стандартной ошибки годового стока между исходными и вычисленными значениями стока для зависимой выборки для всех деревьев от 1 до 7 (табл. 2). Нетрудно видеть, что с ростом числа ветвлений коэффициент детерминации увеличивается, а стандартная ошибка годового стока уменьшается. Для полного дерева коэффициент детерминации между исходными и вычисленными значениями стока для зависимой выборки составляет $R^2 = 0{,}89$, стандартная ошибка стока равна 130 м³/с при стандартном отклонении речного стока 402 м³/с. В то же время распределение стандартной ошибки прогностических значений стока для независимой выборки имеет случайный характер, причем ее оценки довольно сильно варьируют. Как видно из табл. 2, стандартная ошибка для полного дерева имеет минимальную оценку для зависимой выборки и максимальную — для независимой выборки. Результат в определенной степени парадоксальный.

Таблица 2

Статистические оценки годового стока Северной Двины по зависимой и независимой выборкам для всех деревьев решений

Статистическая характеристика	Номер дерева						
	1	2	3	4	5	6	7
Коэффициент детерминации по зависимой выборке	0,89	0,87	0,84	0,78	0,72	0,61	0,39
Стандартная ошибка годового стока Северной Двины по зависимой выборке, м³/с	130,3	140,4	157,5	184,0	207,2	246,8	309,1
Стандартная ошибка годового стока Северной Двины по независимой выборке, м³/с	336,2	287,7	334,2	269,3	261,0	293,1	327,0

Наилучшие оценки прогноза стока Северной Двины получаются для дерева 5, имеющего 4 терминальные вершины и 3 нетерминальные (см. рис. 4). Стандартная ошибка стока Северной Двины по независимой выборке равна 261 м³/с, что составляет 0,65 от величины стандартного отклонения стока (402 м³/с), обычно принимаемой в качестве порогового значения точности долгосрочного прогноза. Отметим, что все

деревья, приведенные в табл. 2, обладают прогностической ценностью. Эти результаты принципиально отличаются от прогноза стока на основе модели МЛР.

На рис. 5 приводится сопоставление фактических и вычисленных по дереву 5 годовых значений стока Северной Двины для зависимой и независимой выборки, которое подтверждает их высокое сходство.

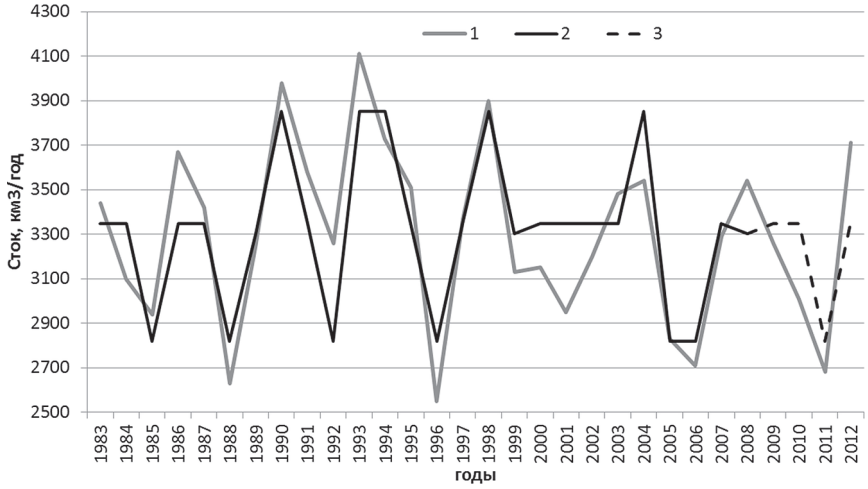

Рис. 5. Сопоставление фактических (*1*) и вычисленных по дереву 5 годовых значений стока Северной Двины (м³/с) для зависимой (1982–2007) (*2*) и независимой (2008–2012) (*3*) выборок

Итак, первые полученные результаты свидетельствуют о широких перспективах использования метода решений деревьев для различных гидрометеорологических задач, включая такую сложную, как долгосрочный прогноз речного стока. Прогноз стока Северной Двины с помощью алгоритма CART обладает высокой точностью, несмотря на то что ветвление происходит формальным путем, и никакой физической основы под ним нет. Очевидно, дальнейшие исследования позволят более четко выявить круг гидрометеорологических задач, для которых может быть использован метод *decision trees*, и решить ряд вопросов методического характера, в частности выбор оптимального размера дерева.

Литература

1. *Андреев И.* Деревья решений — CART: математический аппарат [Электронный ресурс] // Base Group Labs: технологии анализа данных. —URL: Часть 1: https://basegroup.ru/community/articles/math-cart-part1; Часть 2: https://basegroup.ru/community/articles/math-cart-part2.
2. *Малинин В.Н.* Статистические методы анализа гидрометеорологической информации. — СПб.: РГГМУ, 2008. — 407 с.
3. *Чубукова И.А.* Data Mining. — М.: Интернет-университет информационных технологий; Бином, лаборатория знаний, 2008. — 384 с.
4. *Шампандар А.Е.* Деревья классификации и регрессии // Искусственный интеллект в компьютерных играх. — М.: ИД «Вильямс», 2007. — С. 385–401.

5. *Bramer M.* Principles of Data Mining. — Springer, 2007. — 344 p.– DOI:10.1007/978-1-84628-766-4.
6. *Breiman L., Friedman J., Olshen R., Stone C.* Classification and Regression Trees. — Wadsworth, Belmont, CA, 1984. — 358 p.
7. *Classification* and Regression Trees: textbook [Electronic resource]. — Carnegie Mellon University, Statistics Department. — URL: http://www.stat.cmu.edu/~cshalizi/350 /lectures/ 22/lecture-22.pdf
8. *Data* Mining / Википедия-свободная энциклопедия [Электронный ресурс]. — URL: https://ru.wikipedia.org/wiki/Data_mining#cite_note-comp-0
9. *Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R.* Advances in knowledge discovery & data mining. — Cambridge, MA: MIT Press, 1996.
10. *Hunt E.B., Marin J., Stone P.J.* Experiments in induction. — N.Y., Academic Press, 1966.
11. *Hand D.J., Mannila H., Smith P.* Principles of Data Mining. — The MIT Press, 2001. — 546 p.
12. *Interactive* Trees (C&RT, CHAID): Statistica Help / StatSoft inc. [Electronic resource]. — URL: http://documentation.statsoft.com/STATISTICAHelp.aspx?path= Gxx/Indices/InteractiveTreesCRTCHAID_HIndex
13. *Murthy S.* Automatic construction of decision trees from data: A multidisciplinary survey // Data Mining and Knowledge Discovery, 1998, vol. 2, iss. 4, p. 345–389. — DOI:10.1023/A:1009744630224.
14. *Popular* Decision Tree: Classification and Regression Trees (C&RT) / DELL Software [Electronic resource]. — URL: http://documents.software.dell.com/Statistics/ Textbook/Classification-and-Regression-Trees
15. *Pregibon D.* Data Mining // Statistical Computing and Graphics, 1997, vol. 7, p. 8.

Работа выполнена при финансовой поддержке РФФИ (грант 14-05-00837).