Содержание

Предисловие

Данная книга открывает серию пособий, посвященных практическому применению методов машинного обучения на базе популярных статистических пакетов IBM SPSS Statistics и R. В первом выпуске освещается метод деревьев решений. Деревья решений – это эффективный метод машинного обучения, использующийся в прогнозном моделировании. Кроме того, при решении задач бинарной классификации он нередко дополняет метод логистической регрессии. Аналитики кредитного бюро TransUnion для построения скоринговых моделей используют логистическую регрессию, а для отбора переменных в модель логистической регрессии (рассматриваются сотни переменных) – деревья решений CRT или случайный лес. Наша компания при построении прогнозных моделей на основе логистической регрессии использует метод деревьев решений, чтобы сформировать новые переменные для лучшего прогнозирования дефолта. Аналитики Citibank USA разбивают популяцию заемщиков на сегменты, применяя дерево решений, а затем в каждом сегменте строят модели доходности с помощью линейной регрессии или модели риска с помощью логистической регрессии.

Кратко о самой книге. В ней я детально расскажу о том, как строить деревья решений, интерпретировать их, оценивать дискриминирующую способность полученных моделей, улучшать их, сохранять результаты и применять правила классификации/прогноза, полученные с помощью дерева, к новым данным. Кроме того, я расскажу о том, как с помощью дерева решений улучшить модель логистической регрессии. Глава 1 кратко знакомит с терминологией метода деревьев решений, в ней рассказывается о преимуществах и недостатках деревьев, задачах, которые можно выполнить с их помощью. Главы 2–4 посвящены построению деревьев решений в IBM SPSS Statistics 24.0. В главе 2 освещается CHAID – один из самых популярных методов деревьев решений. В главе 3 я покажу, как можно менять параметры дерева CHAID, влияя на результаты классификации. Здесь же я расскажу о том, как можно выполнить биннинг переменных для включения в модель логистической регрессии, использовав дерево решений CHAID и случайный лес. Для иллюстрации выбрана конкурсная задача предсказания отклика ОТП Банка. Кроме того, на данном примере я покажу, как выполняется предварительная подготовка данных и решаются вопросы, связанные с автоматизацией построения моделей (для этого будет использован командный синтаксис SPSS). Код, автоматизирующий процесс построения прогнозных моделей, вы можете в дальнейшем использовать в собственных проектах. В этой же главе будет рассмотрена разработка ансамбля модели логистической регрессии и дерева CHAID. Глава 4 посвящена методам деревьев CRT и QUEST. В главе 5 рассказывается о Редакторе дерева. В главах 6 и 7 я подробно рассмотрю процесс построения и интерпретации деревьев решений в пакетах R CHAID и rpart. Всю необходимую информацию об IBM SPSS Statistics вы найдете на официальном сайте компании IBM. Информацию о программном пакете R можно найти на официальной странице проекта R. Все вопросы, возникшие в ходе чтения книги, можно направлять по адресу info@gewissta.ru.

Освещаемые темы будут интересны маркетологам, риск-аналитикам и другим специалистам, занимающимся разработкой и внедрением прогнозных моделей.

В заключение я хочу поблагодарить моих взыскательных клиентов и коллег из TransUnion, DBS Bank и Citibank USA, в особенности Дмитрия Майорова (Citibank N. A., ArrowModel) и Барри Уилка (Google) за их ценные советы и замечания, высказанные в ходе подготовки книги.

Артем Груздев,
генеральный директор ИЦ «Гевисста»

Глава 1

Введение в метод деревьев решений

1.1. Введение в методологию деревьев решений

Как и регрессионный анализ, деревья решений являются методом изучения статистической взаимосвязи между одной зависимой переменной и несколькими независимыми (предикторными) переменными. Базовое отличие метода деревьев решений от регрессионного анализа заключается в том, что взаимосвязь между значением зависимой переменной и значениями независимых переменных представлена не в виде общего прогнозного уравнения, а в виде древовидной структуры, которую получают с помощью иерархической сегментации данных.

Берется весь обучающий набор данных, называемый **корневым узлом**, и разбивается на два или более **узлов (сегментов)** так, чтобы наблюдения, попавшие в разные узлы, максимально отличались друг от друга по зависимой переменной (например, выделяем два узла с наибольшим и наименьшим процентами «плохих» заемщиков). В роли **правил разбиения**, максимизирующих эти различия, выступают значения независимых переменных (пол, возраст, доход и др.). Качество разбиения оценивается с помощью статистических критериев. Правила и статистики отмечаются на **ветвях** – линиях, которые соединяют разбиваемый узел с узлами, полученными в результате разбиения. Для каждого узла вычисляются **вероятности** в виде **процентных долей** категорий зависимой переменной (если зависимая переменная является категориальной) или средние значения зависимой переменной (если зависимая переменная является количественной). В результате выносится **решение** – спрогнозированная категория зависимой переменной (если зависимая переменная является категориальной) или спрогнозированное среднее значение зависимой переменной (если зависимая переменная является количественной).

Аналогичным образом каждый узел, получившийся в результате разбиения корневого узла, разбивается дальше на узлы, т. е. узлы внутри узла, и т. д. Этот процесс продолжается до тех пор, пока есть возможность разбиения на узлы. Данный процесс сегментации называется **рекурсивным разделением**. Получившаяся иерархическая структура, характеризующая взаимосвязь между значением зависимой переменной и значениями независимых переменных, называется **деревом**.

Иногда для обозначения разбиваемого узла применяется термин **родительский узел**. Новые узлы, получившиеся в результате разбиения, называются **дочерними**

узлами (или **узлами-потомками**). Когда впоследствии дочерний узел разбивается сам, он становится родительским узлом. Окончательные узлы, которые в дальнейшем не разбиваются, называются **терминальными узлами** дерева. Их еще называют **листьями**, потому что в них рост дерева останавливается. Лист представляет собой наилучшее окончательное решение, выдаваемое деревом. Здесь мы определяем группы клиентов, обладающие желаемыми характеристиками (например, тех, кто погасит кредит или откликнется на наше маркетинговое предложение).

Обратите внимание, если вы прогнозируете вероятность значения категориальной зависимой переменной по соответствующим значениям предикторов, дерево решений называют **деревом классификации** (рис. 1.1). Например, дерево классификации строится для вычисления вероятности дефолта у заемщика (на основе спрогнозированной вероятности мы относим его к «плохому» или «хорошему» заемщику). Если дерево решений используется для того, чтобы спрогнозировать среднее значение количественной зависимой переменной по соответствующим значениям предикторов, его называют **деревом регрессии** (рис. 1.2). Например, дерево регрессии строится, чтобы вычислить средний размер вклада у клиента.

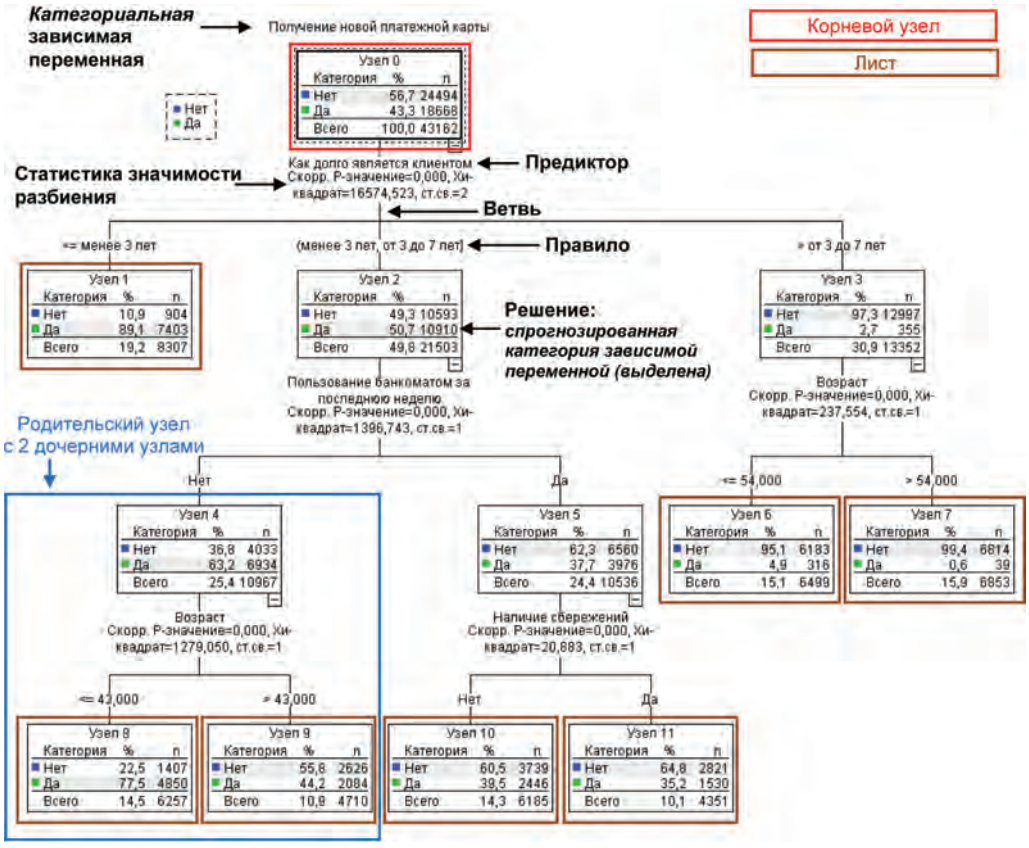

Рис. 1.1 ❖ Пример дерева классификации

Рис. 1.2 ❖ Пример дерева регрессии

1.2. Преимущества и недостатки деревьев решений

Метод деревьев решений обладает рядом преимуществ. Главное из них – это наглядность представления результатов (в виде иерархической структуры дерева). Деревья решений позволяют работать с большим числом независимых переменных. На вход можно подавать все существующие переменные, алгоритм сам выберет наиболее значимые среди них, и только они будут использованы для построения дерева (автоматический отбор предикторов). Деревья решений способны выявлять нелинейные взаимосвязи, сложные взаимодействия, которые нелегко обнаружить в рамках стандартных статистических моделей. Они могут работать с любым типом переменной, таким образом, зависимые и независимые переменные могут быть количественными, порядковыми и номинальными. Деревья решений устойчивы к выбросам, поскольку разбиения основаны на количестве наблюдений внутри диапазонов значений, выбранных для расщепления, а не на абсолютных значениях. Перед построением модели необязательно импутировать пропущенные значения, поскольку деревья используют собственные процедуры обработки пропусков. Требования, выдвигаемые методом деревьев решений к распределению переменных, не являются строгими.

К недостаткам метода деревьев решений можно отнести отсутствие простого общего прогнозного уравнения, выражающего модель (в отличие от регрессионного анализа). Другой недостаток заключается в том, что некоторым методам деревьев решений (например, CRT) свойственно переобучение. Речь идет о ситуации, когда деревья получаются слишком детализированными, имеют много узлов и ветвей, сложны для интерпретации, что требует специальной процедуры отсечения ветвей (она называется прунинг). Наконец, для методов одиночных деревьев характерна проблема множественных сравнений. Перед расщеплением узла дерево сравнивает различные варианты разбиения, число этих вариантов зависит от числа уровней предикторов, как правило, происходит смещение выбора в пользу переменных, у которых большее ко-

личество уровней. Все это обусловливает определенную нестабильность результатов. Небольшие изменения в наборе данных могут приводить к построению совершенно другого дерева. В силу иерархичности дерева изменения в верхних узлах ведут к изменениям во всех узлах, расположенных ниже. Отмечу, что в большей степени вышесказанное относится к методу CRT. Чтобы достигнуть удовлетворительной прогностической способности CRT, один из его разработчиков – Лео Брейман – пришел к идее случайного леса, когда из обучающего набора извлекаются случайные выборки (того же объема, что и исходный обучающий набор) с возвращением, по каждой строится дерево с использованием случайно отобранных предикторов, и затем результаты, полученные по каждому дереву, усредняются. Однако при таком подходе теряется главное преимущество деревьев решений – простота интерпретации.

1.3. Задачи, выполняемые с помощью деревьев решений

Прежде всего деревья решений используются в маркетинге для сегментации клиентской базы. Например, деревья позволяют определить, какие демографические группы имеют максимальный показатель отклика. Эту информацию можно использовать, чтобы максимизировать отклик при будущей прямой рассылке.

Кроме того, деревья применяются для задач прогнозирования и классификации, когда моделируется взаимосвязь между зависимой переменной и предиктором. С этой точки зрения деревья решений сравнивают с логистической регрессией и линейной регрессией. Деревья решений более эффективны, по сравнению с регрессионным анализом, в тех случаях, когда взаимосвязи между предикторами и зависимой переменной являются нелинейными, переменные имеют несимметричные распределения, наблюдаются большое количество коррелирующих между собой переменных, взаимодействие высоких порядков, аномальные значения. Если же предпосылки регрессионного анализа выполняются, то логистическая регрессия (когда зависимая переменная является категориальной) или линейная регрессия (когда зависимая переменная является количественной) может дать лучший результат. Это обусловлено тем, что деревья пытаются описать линейную связь между переменными путем многократных разбиений по предикторам. CHAID делает это за счет расщепления сразу на несколько категорий, CRT и QUEST пытаются уловить эту связь посредством серии бинарных делений, и это может быть менее эффективно, по сравнению с подбором параметров в регрессионном анализе. Однако проблема заключается в том, что данные обычно содержат как линейные, так и нелинейные зависимости, переменные с симметричными и асимметричными распределениями. Поэтому опытный моделер может построить ансамбль логистической регрессии и дерева решений (при условии, что оно использует строгие статистические критерии для отбора предикторов разбиения узлов), чтобы скомпенсировать недостатки обоих методов. Ансамбли дерева решений CHAID и логистической регрессии используются в моделях оттока в телекоме, где данные часто характеризуются переменными с U-образным распределением.

В банковском скоринге деревья решений используются как вспомогательный инструмент при разработке модели логистической регрессии. Приведем конкретные примеры такого применения дерева.

В кредитном скоринге использование нескольких скоринговых карт для одного портфеля обеспечивает лучшее дифференцирование риска, чем использование одной скоринговой карты. Это характерно, когда нам приходится работать с разнородной аудиторией, состоящей из различных групп, и одна и та же скоринговая карта не может работать достаточно эффективно для всех. Например, в скоринге кредитных карточек выделяют сегменты «активные клиенты» и «неактивные клиенты», «клиенты в просрочке» и «клиенты, не имеющие просрочек». Переменные в таких сегментах будут сильно отличаться. Например, для активных кредитных карт утилизация будет сильной переменной, а для неактивных – слабой. И наоборот, может оказаться, что время неактивности для активных клиентов равно 0, а для неактивных клиентов время неактивности окажется сильной переменной. Для этих целей выполняют сегментацию клиентов. Первый способ сегментации – деление на группы на основе опыта и отраслевых знаний с последующей аналитической проверкой. Второй способ – это сегментация с помощью статистических методов типа кластерного анализа или деревьев решений. При этом, по сравнению с кластерным анализом, деревья решений обладают преимуществом: они формулируют четкие правила выделения сегментов, а сами выделенные сегменты статистически значимо отличаются между собой по зависимой переменной. В дальнейшем для каждого из сегментов можно построить собственную модель логистической регрессии, разработать скоринговую карту и сформулировать кредитные правила. В Citibank USA является стандартной практикой делать дерево с двумя-тремя уровнями и в каждом узле подгонять свою модель логистической регрессии. В основе скорингового балла FICO также лежит сегментация на основе деревьев решений. Об эффективности использования сегментации в кредитном скоринге пишет в своей книге «Скоринговые карты для оценки кредитных рисков» известный эксперт по управлению рисками Наим Сиддики[1], а также один из разработчиков алгоритмов скоринга компании FICO Брюс Ходли[2].

С помощью деревьев решений из большого числа предикторов можно выбрать переменные, полезные для построения модели логистической регрессии. Например, из 100 переменных дерево включило в модель 25 переменных, таким образом, у нас появляется информация о том, какие переменные наверняка можно включить в модель логистической регрессии. Метод деревьев решений CRT позволяет вычислить важность переменных, использованных в модели дерева. Мы уже можем ранжировать переменные по степени полезности.

Деревья решений можно использовать для биннинга – перегруппировки категориального предиктора или дискретизации количественного предиктора с целью лучшего описания взаимосвязи с зависимой переменной. Например, при построении модели логистической регрессии часто обнаруживается, что взаимосвязи между количественным предиктором и интересующим событием являются нелинейными. Уравнение логистической регрессии, несмотря на нелинейное преобразование своего выходного значения (логит-преобразование), все равно моделирует линейные зависимости между предикторами и зависимой переменной. Возьмем пример нелинейной зависимости между стажем работы в банке и внутренним мошенничеством.

---

[1] *Наим Сиддики*. Скоринговые карты для оценки кредитных рисков. М.: Манн, Иванов и Фабер, 2014.

[2] *Breiman L.* (2001). Statistical modeling: The two cultures.

Допустим, рассчитанный регрессионный коэффициент в уравнении логистической регрессии получился отрицательным. Это значит, что вероятность совершения внутреннего мошенничества с увеличением стажа работы уменьшается. Однако, выполнив разбивку переменной с помощью дерева CHAID на категории до 12 месяцев, от 12 до 36 месяцев, от 36 до 60 месяцев и больше 60 месяцев, стало видно, что зависимость между стажем и внутренним мошенничеством нелинейная. Первая (до 12 месяцев) и последняя (больше 60 месяцев) категории склонны к внутреннему мошенничеству, а промежуточные сегменты, наоборот, не склонны к внутреннему мошенничеству. После правильной разбивки переменной, проведенной с помощью дерева, связь между предиктором и зависимой переменной становится больше похожа на реальную.

Строя модель логистической регрессии, нередко приходится работать с предикторами, у которых большое количество категорий. Как правило, речь идет о географических переменных (регион, область регистрации, область фактического пребывания заемщика, область торговой точки, где клиент брал кредит) и переменных, фиксирующих профессию или сферу занятости заемщика. Если включить такие переменные в модель логистической регрессии, то переменная с k категориями будет преобразована в $k - 1$ дамми-переменных, которые станут в модели логистической регрессии статистически незначимыми. Только представьте, сколько будет дамми-переменных, если у вас 4 географические переменные с 89 категориями. Исключение таких переменных из анализа также нерационально, поскольку они могут дать ценную информацию. Поэтому можно выполнить биннинг с целью укрупнения категорий, а можно построить по этим четырем переменным дерево решений. В результате дерево укрупнит категории переменных и скомбинирует переменные так, чтобы полученные комбинации характеристик максимизировали различия по зависимой переменной. Такую переменную, где категориями являются терминальные узлы дерева, можно включить в модель логистической регрессии.

Вопросы к главе 1

1. Терминальный узел – это:
   a) самый верхний узел, представляющий всю выборку наблюдений;
   b) узел, в котором рост дерева останавливается;
   c) любой расщепляемый узел;
   d) новый узел, появившийся в результате расщепления узла.

2. Деревья классификации строятся:
   a) для количественной зависимой переменной;
   b) для категориальной зависимой переменной;
   c) для порядковой зависимой переменной;
   d) для номинальной зависимой переменной;
   e) для бинарной зависимой переменной;
   f) для любой зависимой переменной.

3. Деревья регрессии строятся:
   a) для количественной зависимой переменной;
   b) для категориальной зависимой переменной;

c) для номинальной зависимой переменной;
d) для бинарной зависимой переменной;
e) для любой зависимой переменной.

4. В качестве правил разбиения используются:
   a) значения независимых переменных;
   b) значения зависимых переменных;
   c) значения категориальных переменных;
   d) значения количественных переменных.

ЧАСТЬ I

ПОСТРОЕНИЕ ДЕРЕВЬЕВ РЕШЕНИЙ В IBM SPSS STATISTICS

Глава 2

Основы прогнозного моделирования с помощью деревьев решений CHAID

2.1. Запуск процедуры Деревья классификации

Дерево решений в IBM SPSS Statistics можно построить с помощью процедуры **Деревья классификации**. Для вызова процедуры **Деревья классификации** необходимо в меню **Анализ** выбрать **Классификация… Деревья классификации**.

Вы оказываетесь в главном диалоговом окне **Дерево решений** (рис. 2.1). В поле **Зависимая переменная** необходимо перенести одну зависимую переменную. Кнопка **Категории** позволяет включить/исключить из анализа категории зависимой переменной или задать их как целевые. Указание одной или нескольких категорий как целевых не влияет на модель дерева, оценки рисков и результаты классификации. В поле **Независимые переменные** необходимо перенести одну или несколько независимых переменных.

Параметр **Первая переменная принудительно** позволяет задать первую переменную из списка независимых переменных как первую переменную расщепления.

Поле **Переменная влияния** позволяет указать переменную, которая будет определять, насколько большое влияние данное наблюдение оказывает на процесс построения дерева. Наблюдения с более низкими значениями переменной влияния будут иметь меньшее влияние, а наблюдения с более высокими значениями – большее влияние. При этом значения переменной влияния должны быть положительными.

Выпадающий список **Метод построения** позволяет выбрать метод построения дерева.

В правой части окна находятся пять кнопок, используемых для настройки процедуры **Деревья классификации**.

Кнопка **Вывод** задает появление дерева решений и генерацию таблиц. Можно запросить дополнительную статистическую информацию о модели, графическую интерпретацию соответствующих статистик, также можно запросить генерацию правил классификации для модели в SPSS синтаксисе, в SQL или в обычном текстовом формате.

Кнопка **Проверка** позволяет построить модель, отобрав только часть данных, и затем посмотреть, как она работает на оставшейся части, которая была исключена при построении модели.

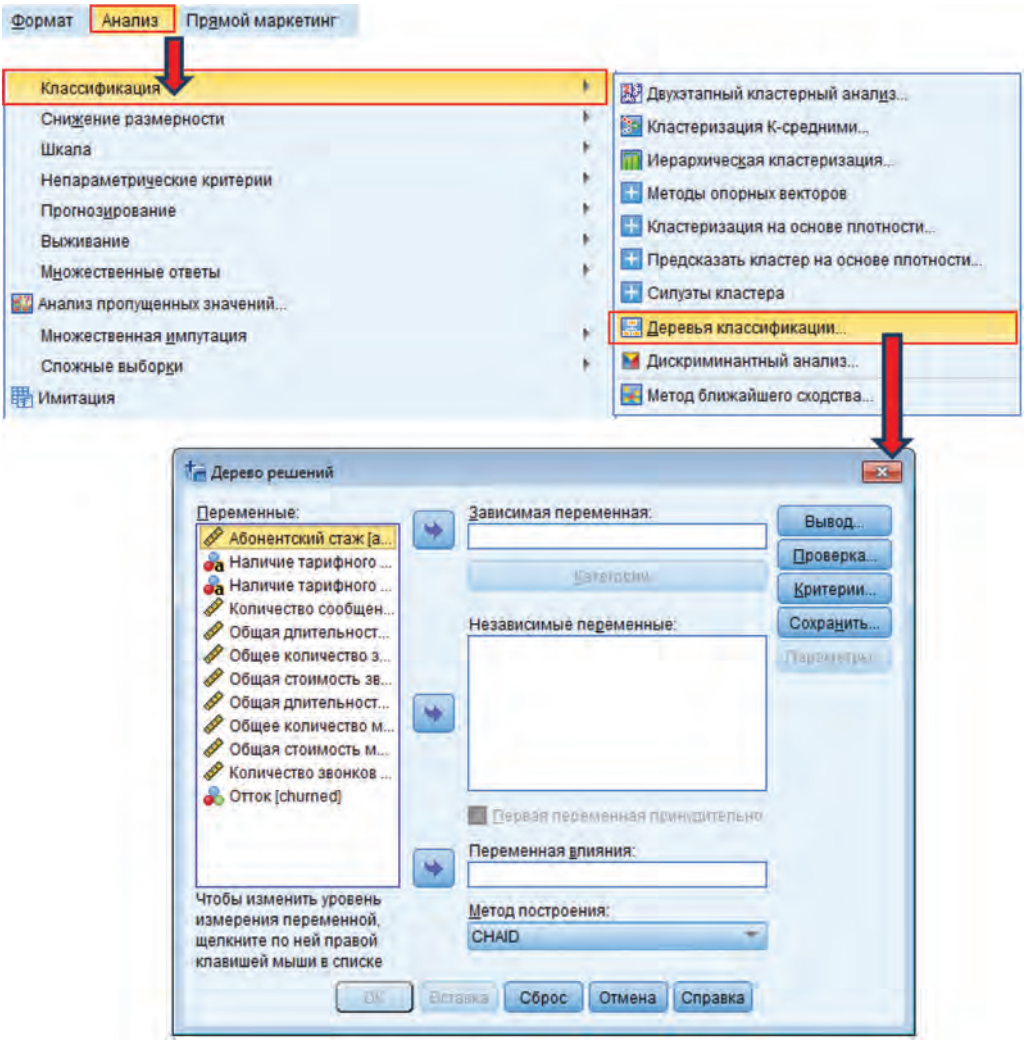

Рис. 2.1 ❖ Запуск процедуры **Деревья классификации**

Кнопка **Критерии** задает значения, которые используются в построении модели, такие как минимальное количество наблюдений в каждой группе или сегменте и уровень значимости, используемый в статистических тестах.

Кнопка **Сохранить** добавляет в активный набор данных переменные результатов анализа для каждого наблюдения:

- номер узла, к которому относится наблюдение;
- спрогнозированное значение зависимой переменной (для количественной зависимой переменной сохраняется спрогнозированное среднее значение, для категориальной зависимой переменной – спрогнозированная категория);
- спрогнозированные вероятности категорий зависимой переменной (только для категориальной зависимой переменной);
- принадлежность к обучающей или контрольной выборке.

Кнопка **Параметры** позволяет задать стоимости ошибочной классификации, априорные вероятности, прибыль и затраты по результатам классификации.

2.2. Четыре метода деревьев решений

Выпадающий список **Метод построения** в диалоговом окне **Дерево решений** (рис. 2.2) позволяет вам выбрать четыре метода деревьев решений: **CHAID** (используется по умолчанию), **Исчерпывающий CHAID**, **CRT**, **QUEST**, **CHAID** (расшифровывается как *Chi-square Automatic Interaction Detector – Автоматический обнаружитель взаимодействий*) – используется процедурой **Деревья классификации** по умолчанию. Он был разработан Гордоном Каасом в 1980 году и представляет собой метод на основе дерева решений, который исследует взаимосвязь между предикторами и зависимой переменной с помощью статистических тестов.

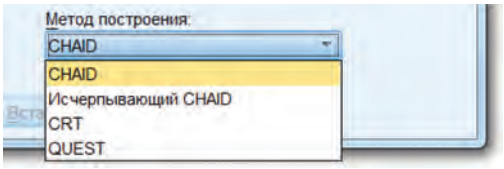

Рис. 2.2 ❖ Выпадающий список **Метод построения**

Каждый раз для разбиения узла выбирается предиктор, сильнее всего взаимодействующий с зависимой переменной. При этом категории каждого предиктора объединяются, если они не имеют между собой статистически значимых отличий по отношению к зависимой переменной, остальные категории рассматриваются как отдельные. Для количественной зависимой переменной используется F-тест, для категориальной зависимой переменной – хи-квадрат Пирсона или хи-квадрат отношения правдоподобия.

Зависимая переменная и предикторы могут быть измерены в номинальной, порядковой и количественной шкале, при этом количественные предикторы преобразовываются в порядковые переменные. CHAID позволяет осуществлять многомерные расщепления узлов. Каждый узел при разбиении может иметь более 2 потомков, поэтому CHAID имеет тенденцию выращивать более раскидистые деревья, чем бинарные методы. Вместе с тем из-за жестких статистических критериев расщепления нередко дерево CHAID получается нереалистично коротким и тривиальным («грубое» дерево), поэтому требуется тонкая настройка уровней значимости для объединения категорий и разбиения узлов. По сравнению с другими методами, CHAID характеризуется умеренным временем вычислений.

Помимо прочего, метод CHAID обладает собственным способом обработки пропущенных значений. Пропуски рассматриваются как отдельная фактическая категория. В ряде случаев это имеет смысл. Например, отказ отвечать на вопрос о доходе или занятости может оказаться предсказательной категорией для зависимой переменной.

Исчерпывающий CHAID является модификацией метода CHAID, предложенной Биггсом, де Виллем и Суеном в 1991 году. Он был разработан для устранения недостатка CHAID – ограниченного набора расщеплений для предиктора.

CHAID прекращает объединение категорий, когда обнаруживает, что все оставшиеся категории статистически различаются между собой. Исчерпывающий CHAID исправляет это, продолжая объединять категории предиктора до тех пор, пока не останутся

только две суперкатегории. Таким образом, он позволяет найти наилучшее расщепление для каждого предиктора и затем выбрать, какой предиктор нужно расщепить.

Исчерпывающий CHAID идентичен CHAID с точки зрения используемых зависимой переменной и предикторов, статистических тестов значимости взаимодействия и способа обработки пропущенных значений. Вместе с тем, поскольку объединение категорий осуществляется более тщательно, чем в методе CHAID, исчерпывающий CHAID требует большего времени вычислений. Надежность результатов исчерпывающего CHAID выше, чем у CHAID.

CRT (расшифровывается как Classification and Regression Tree – Деревья классификации и регрессии) был разработан в 1974–1984 годах профессорами статистики Лео Брейманом (Калифорнийский университет в Беркли), Джеромом Фридманом (Стэнфордский университет), Ричардом Олшеном (Калифорнийский университет в Беркли) и Чарльзом Стоуном (Стэнфордский университет).

Для построения дерева метод CRT использует принцип уменьшения неоднородности в узле. Расщепление узла происходит так, чтобы узел-потомок был более однородным, чем его узел-родитель. В абсолютно однородном узле все наблюдения имеют одно и то же значение целевой переменной (все объекты принадлежат к одной и той же категории целевой переменной). Такой узел еще называют «чистым».

Зависимая переменная может быть измерена в номинальной, порядковой и количественной шкале. Предикторы могут быть измерены в номинальной, порядковой и количественной шкале (подробнее о типах шкал читайте в разделе 2.3 «*Шкалы переменных*»). CRT позволяет только одномерные расщепления узлов. Каждый узел при разбиении может иметь лишь 2 потомков. Поэтому CRT имеет тенденцию выращивать высокие деревья с большим количеством уровней. Часто деревья CRT получаются слишком детализированными, имеют много узлов и ветвей, сложны для интерпретации, при этом усложнение дерева не приводит к повышению прогностической способности дерева (эффект переобучения). Для упрощения структуры дерева и устранения переобучения в методе CRT предусмотрена возможность отсечения ветвей (прунинг). Прунинг позволяет получить дерево «подходящего размера», избежать построения ветвистых, усложненных деревьев и при этом достичь наиболее точной оценки классификации.

Для обработки наблюдений, у которых пропущено значение в предикторе, используются суррогаты – другие предикторы, имеющие сильную корреляцию с исходной независимой переменной. Таким образом, разбиение, задаваемое суррогатом, будет наиболее близко к разбиению, задаваемому исходным предиктором, по которому имеются пропуски. Метод CRT требует большего времени вычислений, по сравнению с другими методами.

QUEST (расшифровывается как *Quick, Unbiased, Efficient Statistical Tree – Быстрое, несмещенное, эффективное статистическое дерево*) был предложен в 1997 году профессорами статистики Вэй Ин Ло (Университет Висконсина-Мэдисона) и Ю Шан Ши (Национальный университет Чун Чен, Тайвань).

Метод QUEST строит дерево следующим образом: для отбора предикторов используются статистические тесты значимости взаимодействия между зависимой переменной и предиктором, а разбиение узлов задается путем выполнения квадратичного дискриминантного анализа с использованием отобранного предиктора. Зависимая переменная может быть измерена только в номинальной шкале. Предикторы могут быть измерены в номинальной, порядковой и количественной шкале.

QUEST имеет схожие с CRT характеристики:

- позволяет только одномерные расщепления узлов;
- каждый узел при разбиении может иметь лишь 2 потомков;
- есть возможность отсечения ветвей (прунинг);
- для обработки наблюдений, у которых пропущено значение в предикторе, используются суррогаты – другие предикторы, имеющие сильную корреляцию с исходной независимой переменной.

Ниже на рис. 2.3 приводится таблица сходств и различий между четырьмя методами деревьев решений, предлагаемых процедурой **Деревья классификации**.

Характеристика метода	CHAID	Exhaustive CHAID	CRT	QUEST
Категориальная зависимая переменная	Да	Да	Да	Да, только номинальная
Категориальные предикторы	Да	Да	Да	Да
Количественная зависимая переменная	Да	Да	Да	Нет
Количественные предикторы	Да, преобразуются в порядковые	Да, преобразуются в порядковые	Да	Да
Тип разбиения	Множественный	Множественный	Бинарный	Бинарный
Цены ошибочной классификации (Построение дерева)	Нет	Нет	Да	Да
Статистические тесты (Отбор предикторов)	Да	Да	Нет	Да
Статистические тесты (Разбиение)	Да	Да	Нет	Нет
Время вычислений	Умеренное	Умеренное	Большое	Умеренное/Большое
Использование априорных вероятностей	Нет	Нет	Да	Да
Пропущенные значения в предикторах	Да, как категория	Да, как категория	Нет, для разбиения используется заменитель	Нет, для разбиения используется заменитель

Рис. 2.3 ❖ Четыре метода деревьев решений

2.3. Шкалы переменных

В зависимости от шкалы (уровня измерения) зависимой переменной и независимых переменных деревья решений применяют различные критерии для отбора предикторов и разбиения узлов. Поэтому важно задать правильную шкалу переменной. В IBM SPSS Statistics существуют три типа шкалы: количественная, номинальная, порядковая.

Количественная шкала содержит информацию о расстояниях между уровнями переменной, порядке уровней и количестве объектов в уровнях. Пример предиктора с количественной шкалой – переменная *Возраст*. Например, я знаю, что расстояние между 25 и 30 в два раза меньше, чем расстояние между 30 и 40, 30-летний на 5 лет старше 25-летнего. Я могу упорядочить уровни по нарастанию или убыванию интенсивности определенного признака (например, по увеличению возраста): после 25 сле-