АЛЕКСЕЙ БОРОВСКИЙ,
кандидат социологических наук, ассистент кафедры отраслевой социологии Киевского национального университета имени Тараса Шевченко

СЕРГЕЙ ЛИТВИНОВ,
кандидат социологических наук, ассистент кафедры отраслевой социологии Киевского национального университета имени Тараса Шевченко

Специфика применения метода деревьев решений в анализе массива данных на примере сравнительного исследования

Abstract

The article tries to prove heuristic potential of the Decision Tree Method in analyzing data of the comparative sociologic research "The Ukrainians and Russians: Looking at Each Other", which was done on the initiative of the Institute of Russian Studies, in Russia by the company GfK RUS from June 27 until July 11, 2008, in Ukraine — by the company GfK Ukraine from June 19 until July 7, 2008.
The main task of the Decision Tree is to display and visualize a covered categorical data structure, as if it were peculiar to them, to analytical separation of empirical data by statistical methods.
By the use of the Decision Tree Method the authors have succeeded to build portraits of the respondents and discover from the structure of massifs the most dependent variables. Thus the authors conclude that in Ukraine regional and socio-cultural factors are the most important determinant when evaluating relations with Russia and in Russia socio-demographical characteristics of respondents are more significant.

Введение

В последнее время в информационном пространстве Украины и России наблюдается существенное увеличение интереса к проблемам межгосударст-

венного взаимодействия двух стран и роли общественного мнения в этих отношениях. Актуализация этого явления вызвана как нарастанием политических противоречий, так и рядом социальных причин. В книге, посвященной исследованию национально-гражданских идентичностей и толерантности в России и Украине, среди социальных факторов, отмечаются усиливающиеся межэтнические противоречия и продолжающаяся социальная дезориентация людей в условиях ценностно-нормативной неопределенности и аномической деморализованности значительной части населения [Национально-гражданские идентичности, 2007: с.25]. Такое состояние большинства населения, безусловно, должно влиять на оценку в массовом сознании места и роли межгосударственных отношений. Практический интерес для нас представляет и сам анализ общественного мнения в двух странах, и методологические особенности анализа данных сравнительного социологического исследования, реализованного авторами данной публикации.

Общественное мнение в определенной мере детерминируют постоянно появляющиеся сообщения украинских и российских социологических центров, как правило, представляющие количественный анализ одномерных и двумерных распределений ответов респондентов на вопросы. Такие результаты, однако, не дают целостной картины соотношения между двумя объектами сравнения. Незамеченным остается огромное число социальных факторов, в первую очередь социокультурных различий и национальных особенностей формирования общественного мнения.

В таких условиях практическое значение приобретает метод анализа, используемый при классификации большого числа неоднородных социальных данных. В предлагаемой статье предпринята попытка обоснования эвристического потенциала **метода "деревьев классификации" (или "деревьев решений")** в анализе массива данных сравнительного социологического исследования "Украинцы и россияне: взгляд друг на друга", которое было осуществлено в 2008 году по инициативе Института изучения России.

Что такое "деревья классификации"?

"Деревья классификации" (classification trees) — сравнительно молодой метод data meaning, одна из эвристических процедур глубинного анализа данных. Первые шаги в этом направлении были сделаны в конце 50-х годов XX века Ховлендом и Хантом.Основополагающей для метода деревьев считается более поздняя работа Ханта, Мэрина и Стоуна "Индуктивные эксперименты" (Experiments in Induction), опубликованная в 1966 году. Уже в 1980-х и особенно в 1990-х алгоритмы деревьев классификации стали популярным инструментом биологических и медицинских исследований, а также языком моделирования процесса принятия решений в науках об управлении (см.: [Деревья классификации, s.a.]). Реализующие метод программные продукты в наше время прочно вошли в набор средств "добычи данных".

Иногда *classification trees* относят к алгоритмам так называемого интеллектуального анализа, что подразумевает диалоговый режим и автоматизацию процесса поиска оптимального решения (см.: [Classification, s.a.]). В компетенции пользователя остается корректная формулировка задачи, выбор наиболее адекватных статистических критериев, контроль процесса автоматизированной обработки и интерпретация полученных результатов.

Метод "деревьев классификации" сочетает в себе преимущества алгоритмов, реализованных на современной вычислительной технике, с творческим участием человека в подготовке исходных данных, формулировке гипотез, в теоретическом осмыслении продукта автоматизированной классификации — графа ("дерева") решения. В этой особенности заключаются как плюсы, так и минусы. К первым следует отнести гибкость метода в отношении исходных данных, возможность использования разных статистических критериев для классификации, наглядность и хорошая интерпретируемость дерева решения. Ко вторым — статистическая "слабость" результата, отсутствие критериев надежности классификации данных, функция распределения которых была бы изучена и табулирована. Поэтому метод деревьев классификации следует считать разведочным. Он не может быть использован в традиционном конфирматорном подходе к доказательству статистических гипотез. Напротив, результаты автоматической классификации облегчают формулировку таковых. Однако круг задач данного метода намного шире его сугубо технологического применения. Главная задача дерева решения — проявить и визуализировать скрытую категориальную структуру данных, как бы присущую им самим по себе, до аналитического расчленения статистическим скальпелем. Поэтому корректное использование *classification trees* позволяет не только сэкономить массу времени и ресурсов, но и достигнуть качественно иного уровня объяснения эмпирических зависимостей (см.: [Берестнева, Муратова, 2004]). Ступенчатая классификация объектов по множеству переменных-предикторов, регрессия зависимой переменной, формулировка количественных условий отбора объекта в одну из заранее выделенных групп по наблюдаемым значениям тестовых переменных — вот далеко не полный перечень приложений метода деревьев. Что касается сути и разновидностей метода, то мы отсылаем читателя к соответствующей методологической литературе (см., напр.: [Эффективная сегментация, s.a.; Classification, s.a.; Tsien, Fraser et al., s.a.]).

Основная идея метода

Остановимся кратко на главной идее деревьев решений. Она заключается в следующем. Пусть задано множество из признаков, квантифицированных числовыми переменными — интервальными, порядковыми или номинальными. Один из этих признаков (он должен быть категориальным) мы рассматриваем как зависимый, а остальные $n - 1$ — как предикторы вариации его значений. Взаимное влияние независимых переменных друг на друга нас не интересует. Мы хотим выделить переменную, которая позволяет наилучшим образом сгруппировать объекты, различающиеся по зависимой переменной (целевому параметру). Иными словами, найти переменную, группировка по которой обеспечивает возможность выделения подмножеств объектов, которые будут максимально отличаться по вариации целевого параметра внутри подмножеств. Найдя такую переменную, мы рассматриваем полученные k_1 подмножеств в качестве целевых параметров второго уровня, а оставшиеся $n - 2$ независимых переменных — как предикторы целевых параметров. Затем процедура повторяется в каждом из k_1 случаев. Мы получаем k_2 целевых параметров третьего уровня и т.д. k_i параметров $i + 1$-го уровня. Граф подмножеств ветвится до тех пор, пока группы

выделяемых объектов не станут слишком малы или не будут исчерпаны все $n - 1$ исходных предикторов. При этом на дереве отражаются только те группы объектов (и классификационные переменные), которые значимо различаются по вариации зависимого признака. Соответственно, те "ветки" i-го уровня, которые не удается разделить на значимо отличающиеся подмножества ни одним из $n - i$ предикторов, становятся тупиковыми.

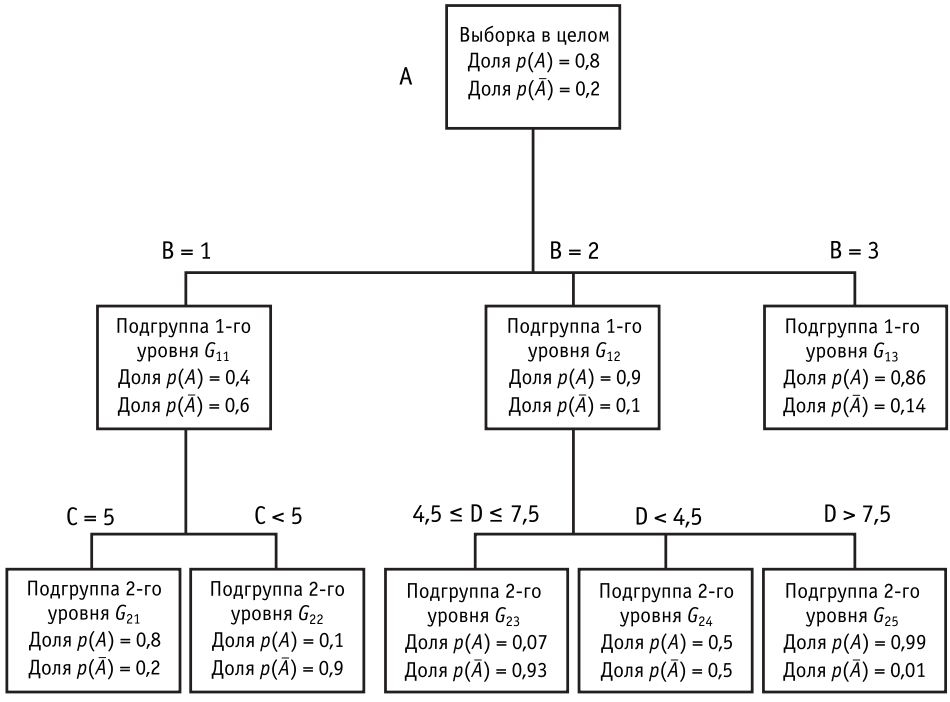


Рис. 1. Условный пример дерева решения, построенного по методу CHAID

На приведенном графе зависимая переменная A — дихотомическая. В теории деревьев решений она носит название целевой переменной (параметра) или метки класса. Целевая переменная является вершиной ветвления. Выделяемые алгоритмом подгруппы образуют узлы графа. Узлы выделяются на основании условия (правила) отбора значений независимой переменной атрибута. Так, в дереве решения на рисунке 1 узлы первого порядка выделяются атрибутом B по правилу: если $B(j) = 1$, то объект j принадлежит подмножеству G_{11}; если $B(j) = 2$ — подмножеству G_{12}; если $B(j) = 3$, объект относится к подмножеству G_{13}.Подмножества первого узла полностью исчерпывают собой исходное множество (выборку) $G(A) = G_{11} \cup G_{12} \cup G_{13}$, а правило описывает все значения атрибута B. Номинальная переменная B принимает значения 1, 2, 3; порядковая C — целые значения от 1 до 5; интервальная переменная D изменяется в границах от 0 до 10. Подобранное алгоритмом правило разбивает область значений D на три непересекающихся интервала. Ответвления узлов первого уровня образуют узлы второго уровня, так что дочерние подмножества, образованные по специфическому для каждого узла первого порядка правилу, полностью исчерпывают собой исходное множество. На рисунке 1 ветка G_{13} тупиковая, а узлы G_{11} и G_{12} рас-

щепляются на узлы второго уровня по правилу для атрибута C (G_{11}) и атрибута D (G_{12}). Конечные узлы дерева G_{13}, G_{21}, G_{22}, G_{23}, G_{24}, G_{25} называются узлами решения, или более поэтично — листами. Интерпретация дерева решения в данном условном примере также достаточно проста. Наибольшей "разделяющей силой" относительно доли признака A обладает атрибут B, образующий узлы первого порядка. Поэтому его можно считать наиболее значимым для вариации A. Иначе говоря, распределение A больше зависит от B, чем от C или D, точнее, зависит от B в первую очередь (о "влиянии" здесь можно говорить только в нестрогом и очень широком смысле слова). Кроме того, по графу легко выделить различные подгруппы с отличающимися средними долями целевого параметра. Максимум $p(A) = 0{,}99$ можно наблюдать в подгруппе G_{25}, выделяемой правилом "$B = 2$ и $D > 7{,}5$". Абсолютный минимум $p(A) = 0{,}99$ достигается в подгруппе G_{23} (правило "$B = 2$ и $4{,}5 \le D \le 7{,}5$"). Следует также обратить внимание на минимум $p(A) = 0{,}1$ в листе G_{22} ("$B = 1$ и $C < 5$"). Остальные листы можно упорядочить по метке класса между минимумом и максимумом.

Что касается качества полученного дерева решения, то оно имеет две составляющие — точность и надежность. Точность классификации естественным образом можно оценить с помощью процента правильно классифицированных объектов. В отдельных случаях, например при анализе медицинских данных, важность правильной классификации неодинакова для различных узлов. Для учета этих различий используют понятия априорной вероятности и цены ошибки классификации [Деревья классификации, s.a.]). Мы не будем их рассматривать; заметим только, что если выбрать пропорциональные величине классов априорные вероятности, а цену ошибки для всех классов считать одинаковой, то мерой качества классификации будет доля правильно классифицированных объектов. Вторую составляющую качества решения, надежность, оценить гораздо сложнее. Статистических критериев для этого просто не существует. В работе [Ростовцев, s.a.] предлагается использовать бутсреп, методики размножения исходной выборки, чтобы с помощью вычислительных процедур, а не предельных аппроксимаций проверить устойчивость древовидной классификации, а следовательно ее надежность.

Область применения, требования и возможности "дерева классификации"

Как использовать результаты анализа дерева решения? Основных применений три.

1. Описание данных. Полученный граф удобно использовать вместо многих таблиц для наглядного представления структуры данных.
2. Классификация объектов и построение иерархии переменных-критериев классификации. Удобство дерева решения для этой цели очевидно.
3. Если метка класса континуумальна, деревья решений позволяют установить зависимость целевой переменной от независимых предикторов. К этому классу относятся задачи численного предсказания значений целевой переменной (регрессия).

Очевидно, сфера применения метода "деревьев классификации" пересекается с методами дискриминантного анализа (если целевая переменная дихотомическая), кластерного анализа, дисперсионного и порядкового регрессионного анализа. Но его преимущество кроме большей наглядности состоит еще и в возможности одновременного решения нескольких задач с помощью одного дерева. Кроме того, метод предусматривает меньшую формализацию и конкретизацию начальных условий, что делает его более гибким и привлекательным для практического использования. Те же преимущества обеспечивают перспективность "деревьев решений" как инструмента социологического анализа анкетных данных [Ростовцев, s.a.; Толстова, 2000; Украинское общество, 2007].

Сжатое описание данных, как и построение эмпирической классификации, относится к важнейшим проблемам обработки данных, если данные представляют собой набор множества переменных разного уровня квантификации, зависимости и отношения между ними *a priori* не определены. Поэтому на первом этапе обработки — до выдвижения статистических гипотез — уместна разведывательная стратегия анализа. Одной из возможных ее реализаций является применение группы методов classification tree. Некоторые авторы рекомендуют использовать "деревья решений" там, где необходимо получить однозначные рекомендации на основании эмпирически вычисленных правил, например, для выдачи кредитов, оперативной диагностики больных и т.д. [Национально-гражданские идентичности, 2007; Classification, s.a.].

Базовые алгоритмы

На сегодняшний день предложен целый ряд критериев, с помощью которых можно оценить значимость различий, а также алгоритмов построения графа классификации. Существует значительное число алгоритмов, реализующих "деревья решений", например, NewId, ITrule, CN2 и т.д. Но наибольшее распространение получили следующие алгоритмы (см.: [Деревья классификации, s.a.; Деревья решений, s.a.; Эффективная сегментация, s.a.]):

- **CHAID (CHi-squared Automatic Interaction Detector)** — разработчик Г.В.Касс (1980). "Метод построения деревьев решений, в котором для получения оптимального разбиения используется критерий связи между категориальными переменными χ^2 (в случае, если целевая переменная является количественной, используется F-критерий). Исходно целевая переменная и переменные-предикторы могут быть как количественными, так и категориальными, однако количественные предикторы при построении дерева преобразуются в категориальные (количеством категорий можно управлять)" [Толстова, 2000: с.]. Реже используются алгоритмы FACT (Loh & Vanichestakul, 1988), THAID (Morgan & Messenger, 1973) или AID (Morgan & Sonquist, 1963).
- **Exhaustive CHAID (Исчерпывающий CHAID)**. Модификация метода CHAID. "Его преимуществом является то, что в процессе построения дерева анализируется большее количество возможных разбиений, а недостатком — более медленная скорость работы. Этот метод накладывает на типы целевой переменной и предикторов те же ограничения, что и метод CHAID" [Эффективная сегментация, s.a.].

- **C&RT (Classification And Regression Trees)**, дословно — метод классификации и построения деревьев регрессии предложен Л.Брейманом и др. (1984). В отличие от двух описанных выше методов основан не на статистических критериях, а на уменьшении неоднородности подгрупп (узлов). Для анализа могут быть использованы как количественные, так и категориальные целевая переменная и переменные-предикторы. Наилучший результат достигается в том случае, если все переменные в анализе являются количественными.
- **QUEST (Quick, Unbiased, Efficient Statistical Trees)**, то есть "быстрые, несмещенные, эффективные статистические деревья" (Loh & Shih, 1997). В данном методе для выбора предикторов применяются различные критерии, в зависимости от типа потенциального предиктора. Метод позволяет избегать смещений, связанных с выбором предикторов с большим количеством категорий. Целевая переменная в данном случае должна быть категориальной. Переменные-предикторы могут быть как количественными, так и категориальными.
- **C4.5.** Разработчик — Р.Квинлан (1993). Алгоритм построения дерева решений, в котором количество ветвлений узла не ограничено. Не предназначен к работе с непрерывным целевым полем, поэтому решает только задачи классификации.

Особенность всех названных алгоритмов, которая определяет специфику метода деревьев решений, состоит в том, что если один раз был выбран атрибут, по которому было произведено разбиение на подмножества, то алгоритм не позволяет вернуться назад и выбрать другой атрибут, который дал бы лучшее разбиение. Поэтому на этапе построения нельзя сказать, даст ли выбранный атрибут оптимальное разбиение.

Пример применения анализа деревьев к выделению критериев эмпирической классификации респондентов

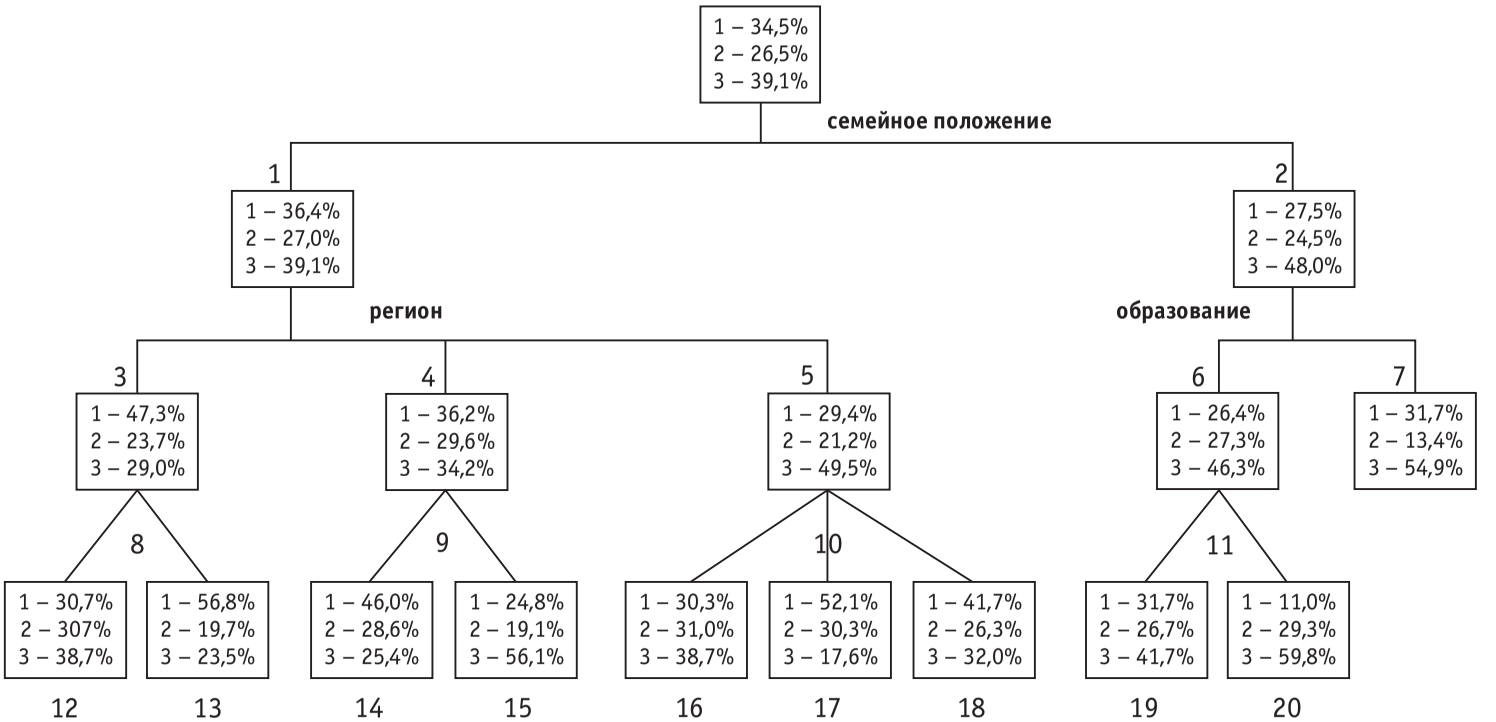
Наш опыт применения алгоритма Tree Analysys в SPSS 13.0 показывает высокую эффективность метода деревьев классификации в обработке сложных массивов данных социологических исследований. Метод был реализован нами в ходе обработки данных второй волны сравнительного исследования "Украинцы и россияне: взгляд друг на друга", проведенного по заказу Института изучения России. В России опрос проводился компанией "GfK RUS" с 27 июня по 11 июля 2008 года, в Украине — компанией "GfK Ukraine" с 19 июня по 7 июля 2008 года.

Опрос респондентов проводился методом личного интервью по месту жительства. Целью опроса было выявление наиболее приближенных оценок состояния межгосударственных отношений между двумя странами. Выборочная совокупность построена по схеме многоступенчатой выборки, полученной методом случайного отбора (в России — 2196 интервью, в Украине — 1313). Теоретическая статистическая погрешность выборочной оценки доли биноминального признака с распределением 50% : 50% при доверительной вероятности $p = 0{,}95$ для Украины не превышает 2,7%, для России — 2,1%.

Одним из главных заданий было вычисление интегрального индекса добрососедства (ИД). ИД строился на основе еще 6 индексов: трех простых — базового индекса отношений (БИО), индекса динамики отношений между странами (ИДОС), индекса динамики отношений между народами (ИДОН) и одного сложного — индекса интереса к другой стране, ее политической, экономической и культурной жизни (ИИД). При этом нам было важно понять: 1) от каких именно факторов в наибольшей мере зависит ИД и 2) какие группы респондентов характеризуются полярными значениями индекса. Использовать для этого описательную статистику и проверку гипотез для построения классификации было бы неэффективно, так как в массиве одних только социально-демографических переменных насчитывалось 9. Если даже сгруппировать данные, то на базе 9 переменных получается не менее $2^9 = 512$ градаций. К тому же большая часть этих градаций при реализованных объемах выборки были бы недостаточно наполненными. Что касается гипотез о влиянии, то мы не считали себя достаточно компетентными для их исчерпывающей формулировки. Первая часть задачи могла быть решена с помощью логистической регрессии. Но в массиве были как категориальные, так и количественные переменные, которые могли рассматриваться в качестве потенциальных предикторов индекса добрососедства. Кроме того параллельно необходимо было решить задачу построения классификации, выделения критериев, по которым различаются группы респондентов с высоким и с низким показателем ИД. Указанным требованиям удовлетворяла методика CHAID алгоритма деревьев решений в SPSS 13.0. Перед тем, как применить методику к нашим данным, мы построили зависимую переменную, значения которой были вычислены как факторные значения (*factor scores*). В качестве индикаторов для факторного анализа мы отобрали переменные индексов БИО, ИДОС, ИДОН и ИИД. Оказалось, что наилучшим образом вариацию индикаторов описывает двухфакторная модель, в которой в первый фактор вошли БИО, ИДОС и ИДОН, а во второй — три переменные-компоненты ИИД. Оценка индекса добрососедства производилась на основе первого фактора, отражающего базовый уровень оценки респондентами украино-российских отношений. В соответствии с методикой расчета факторных значений с помощью регрессии была рассчитана итоговая непрерывная переменная с нормальным распределением значений от –3 до 3. Перед построением дерева классификации она была преобразована в категориальную путем разбиения на терцильные интервалы. Значение "1" шкалы соответствует нижнему терцилю (плохие отношения между государствами), "2" — среднему терцилю (нейтральные отношения), а "3" — верхнему терцилю (хорошие отношения). Именно эта сконструированная переменная была взята в качестве зависимой в Tree Analysis. Множество зависимых переменных включало все социально-демографические признаки и переменные, на основе которых рассчитывались индексы ИИД и ИИС (симметричный ИИД индекс интереса к собственной стране, см. приложение). Так как потенциальные предикторы представляли собой переменные разных типов и нам была нужна классификация, где все градации одного и того же предиктора располагались бы на одном уровне ветвления дерева решения, была выбрана методика CHAID. Минимальная наполненность подгрупп была определена в 50 единиц.

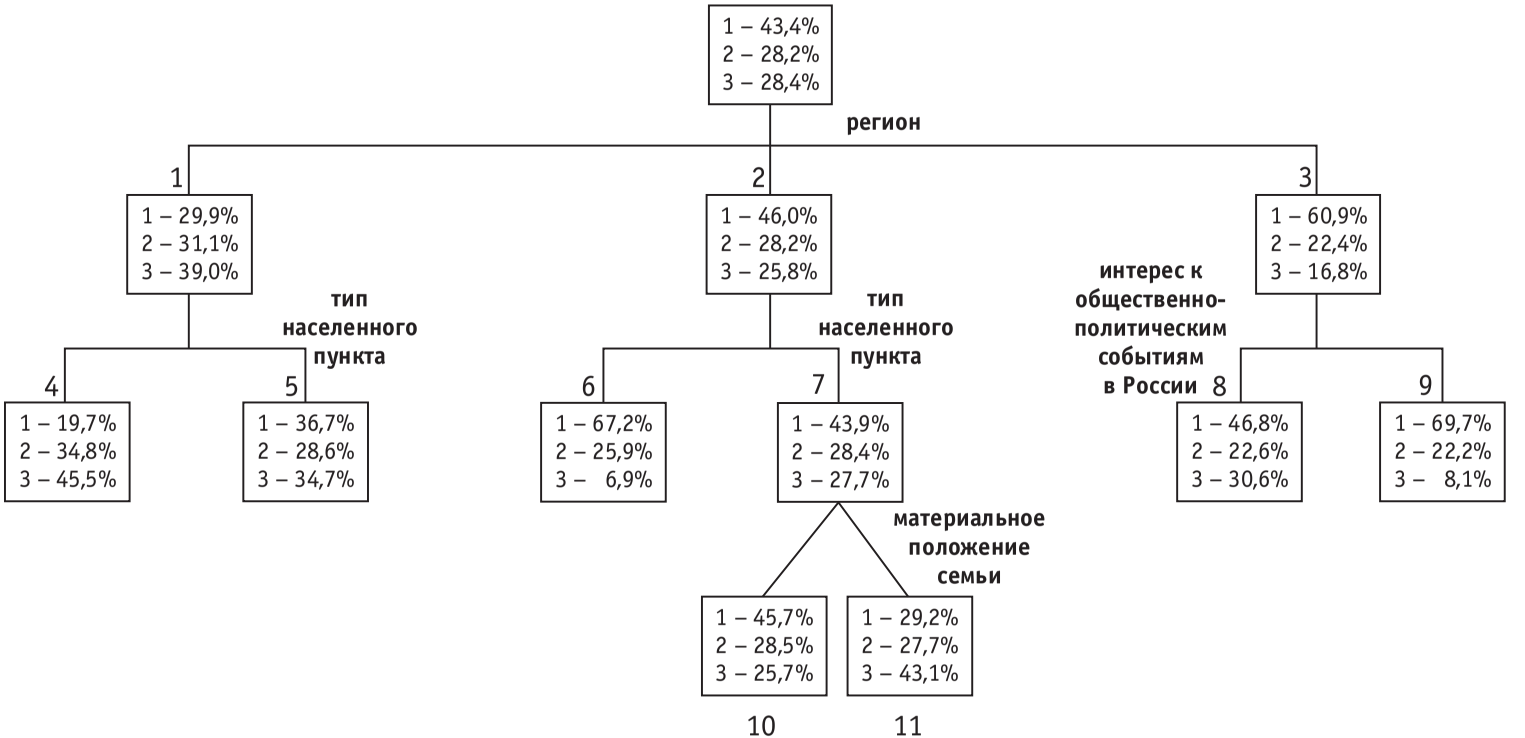
Полученные нами деревья показательны и по украинской, и по российской выборке. Классификационное дерево по российской выборке позволяют правильно классифицировать 71% респондентов (см. рис. 2) , по украинской — 74% (см. рис. 3). Ключевым дифференцирующим российских респондентов признаком является семейное положение, точнее принадлежность к группе холостых (второй узел графа, Node 2). Оценка холостыми респондентами украинско-российских отношений выше, чем в целом по выборке: факторные значения из верхнего терциля встречаются среди них на 9% чаще — у 48% респондентов против 39,1% в общем по массиву. Анализируя социально-демографические характеристики группы холостых можно прийти к выводу, что на самом деле этот признак маркирует возрастные отличия: 82% холостых составляют люди, младше 30 лет. Наиболее оптимистичны среди них две подгруппы: люди с высшим образованием (хорошими считают отношения между Россией и Украиной 55%, а плохими — 32%) и жители села или поселка без высшего образования (60% и 11% соответственно). Большинство же респондентов расслаивается по другим признакам, региональному и поселенческому. "Оптимисты" проживают в Москве и в городах с населением до 100 тыс. Северо-Западного региона, "пессимисты" в Южном регионе и на Дальнем Востоке. Отсюда можно сделать вывод, что относительно более высокие оценки украино-российских отношений дают респонденты, принадлежащие к благополучным и социально оптимистичным группам, а также, так сказать, "простые" люди с низким показателем социального цинизма.

Если обратиться к анализу украинской выборки, то там ситуация несколько иная. Во-первых, регионализм в Украине не только является главным фактором расслоения оценок отношений между странами, но и выделяет количественно более дифференцированные группы, чем это можно наблюдать на российской выборке. Причем получается парадоксальная вещь: различные по всем социокультурным параметрам Западный и Восточный регионы попали в дереве решений в один узел (Node 2). Оба региона демонстрируют "нормально плохую" оценку отношений между странами. Особенно много "пессимистов" в городах с населением 51–100 тыс. человек, которые представляют собой социально-депрессивные социумы ("пессимистов" на 60% больше, чем "оптимистов"). "Оптимисты" локализованы в селах, поселках и городах с населением до 50 тыс. человек и в городах с населением больше 100 тыс. К "оптимистам" принадлежат респонденты с высоким финансовым благосостоянием, которые не испытывают затруднений в удовлетворении важнейших материальных потребностей. Самая большая разница в пределах Украины наблюдается между оценками респондентов из Южного региона, с одной стороны, и Киева, Северного и Центрального регионов — с другой. В первом случае число "оптимистов" относится к числу "пессимистов" как 1 : 3,6, во втором "пессимистов" больше, чем "оптимистов", в 1,35 раза. "Оптимисты" здесь — это люди, проживающие в малых городах или в Киеве, которые умеренно интересуются культурно-спортивной жизнью в Украине. В то же время к числу крайних "пессимистов" в Южном регионе принадлежат те, кто проявляет повышенный интерес к общественно-политической жизни в России. На наш взгляд, это позволяет говорить о социокультурной детерминации оценок отношений между Украиной и Россией украинскими респондентами. Взгляд "пессимистов" Юга Украины



1 – женат (замужем); живут вместе; разведён; живут отдельно; вдовец (вдова)
2 – холост (не замужем)
3 – Южный регион; Дальний Восток
4 – Поволжье; Центральный регион; Сибирь; Урал
5 – Северо-Западный регион; г. Москва
6 – нет высшего образования
7 – есть высшее образование
8 – среднедушевой доход семьи
9 – тип населенного пункта
10 – тип населенного пункта
11 – тип населенного пункта
12 – до 4000 руб.
13 – 4000 руб. и больше
14 – село, поселок; город с населением 100–499 тыс. жителей
15 – город с населением до 100 тыс. жителей; город с населением свыше 1 млн жителей
16 – село, поселок; город с населением 100 тыс. — 1 млн жителей
17 – город с населением свыше 1 млн жителей
18 – город с населением до 100 тыс. жителей
19 – города всех типов
20 – село, поселок

Рис. 2. Классификационное дерево по оценке украино-российских отношений (Россия)



1 – Киев; Северный и Центральный регионы
2 – Западный и Восточный регионы
3 – Южный регион
4 – город с населением до 50 тыс. жителей; город с населением более 500 тыс. жителей
5 – город с населением 51–499 тыс. жителей
6 – город с населением 51–100 тыс. жителей
7 – город с населением до 50 тыс. жителей; город с населением более 100 тыс. жителей
8 – мало интересуюсь; затрудняюсь ответить; совершенно не интересуюсь
9 – очень интересуюсь; в общем интересуюсь
10 – плохое или среднее материальное положение
11 – отличное материальное положение (“хватает на всё”)

Рис. 3. Классификационное дерево по оценке украино-российских отношений (Украина)

словно обращен в сторону России как референтного (или даже "своего") политического пространства. "Оптимисты" из Центральной и Южной Украины наоборот обращены в сторону собственного культурного пространства. Возможно, оптимистическое восприятие отношений между странами связано именно с аполитичностью представителей данной группы. Таким образом, можно заключить, что для Украины важнейшими детерминантами оценки отношений с Россией является региональный и социокультурный факторы, а в России более значимы социально-демографические характеристики респондентов. Регионализм в Украине значит гораздо больше, чем в России. Эти выводы, как и эмпирический портрет "оптимистов" и "пессимистов" гораздо сложнее было бы получить, применяя традиционные методы анализа двумерных распределений. По нашему мнению, эвристический потенциал метода анализа деревьев проявляется как раз на этапе обобщения и группировки данных, в ходе построения эмпирических типологий. Хотя, безусловно, можно использовать этот подход и для того, чтобы попытаться "увидеть" скрытую структуру данных на этапе выдвижения предварительных гипотез, то есть как вспомогательный эксплораторный инструмент. Социологам еще предстоит немалая работа по анализу эффективности применения метода классификационных деревьев и его конкретных алгоритмов к данным социологических исследований, а также по выяснению оптимальных условий применимости этого интересного метода.

ПРИЛОЖЕНИЕ

Переменные-предикторы, использованные при построении классификационных деревьев для признака "Оценка отношений между Украиной и Россией"

1. ИНТЕРЕСУЕТЕСЬ ЛИ ВЫ СОБЫТИЯМИ, ПРОИСХОДЯЩИМИ В УКРАИНЕ? (общественно-политические события)
2. ИНТЕРЕСУЕТЕСЬ ЛИ ВЫ СОБЫТИЯМИ, ПРОИСХОДЯЩИМИ В УКРАИНЕ? (экономические события)
3. ИНТЕРЕСУЕТЕСЬ ЛИ ВЫ СОБЫТИЯМИ, ПРОИСХОДЯЩИМИ В УКРАИНЕ? (культурно-спортивная жизнь)
4. ИНТЕРЕСУЕТЕСЬ ЛИ ВЫ СОБЫТИЯМИ, ПРОИСХОДЯЩИМИ В РОССИИ? (общественно-политические события)
5. ИНТЕРЕСУЕТЕСЬ ЛИ ВЫ СОБЫТИЯМИ, ПРОИСХОДЯЩИМИ В РОССИИ? (экономические события)
6. ИНТЕРЕСУЕТЕСЬ ЛИ ВЫ СОБЫТИЯМИ, ПРОИСХОДЯЩИМИ В РОССИИ? (культурно-спортивная жизнь)
7. ПОЛ
8. ВОЗРАСТ
9. КАКОЕ УЧЕБНОЕ ЗАВЕДЕНИЕ ВЫ ОКОНЧИЛИ ПОСЛЕДНИМ?
10. КАКОЕ ВЫСКАЗЫВАНИЕ НА ЭТОЙ КАРТОЧКЕ НАИЛУЧШИМ ОБРАЗОМ ОПИСЫВАЕТ ВАС И ВАШУ СЕМЬЮ?

11. КЕМ ВЫ РАБОТАЕТЕ В НАСТОЯЩЕЕ ВРЕМЯ?
12. КАКИМ БЫЛ ДОХОД ВАШЕЙ СЕМЬИ В ПРОШЛОМ МЕСЯЦЕ В РАСЧЕТЕ НА ОДНОГО ЧЛЕНА СЕМЬИ
13. СЕМЕЙНОЕ ПОЛОЖЕНИЕ В НАСТОЯЩЕЕ ВРЕМЯ
14. РЕГИОН
15. РАЗМЕР И ТИП НАСЕЛЕННОГО ПУНКТА
16. КЛАССИФИКАЦИЯ ESOMAR

Литература

Берестнева О.Г., Муратова Е.А. Построение логических моделей с использованием деревьев решений // Известия Томского политехнического университета. — 2004. — Т. 307. — № 2. — С.154–160.

Деревья классификации. — <http://www.statsoft.ru/home/textbook/modules/stclatre.html> (s.a.).

Деревья решений — общие принципы работы. — <http://www.basegroup.ru/library/analysis/tree/description/> (s.a.).

Национально-гражданские идентичности и толерантность. Опыт России и Украины в период трансформации / Под ред. Л.М.Дробижевой, Е.И.Головахи. — К., 2007.

Елманова Н. Построение деревьев решений // Введение в Data Mining. Ч.3. — <http://www.interface.ru/fset.asp?Url = /misc/vvdm_p3.htm&anchor = 2> (s.a.).

Отличия алгоритма дерева решений от ассоциативных правил в задачах классификации. — <http://www.spellabs.ru/DecisionTreesVsAssociationAlgorithm.htm> (s.a.).

Ростовцев П.С. Автоматизация анализа анкетных данных. — <http://nesch.ieie.nsc.ru/13ROST8.html> (s.a.).

Толстова Ю.Н. Анализ социологических данных: Методология, дескриптивная статистика, изучение связей между номинальными признаками. — М., 2000.

Украинское общество в европейском пространстве / Под ред. Е.Головахи, С.Макеева. — К., 2007.

Эффективная сегментация при помощи деревьев решений. — <http://www.spss.com.ua/products/answertree/> (s.a.).

Classification: Basic Concepts, Decision Trees and Model Evaluation. — http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf (s.a.).

Tsien L.C., Fraser S.F.H., Long J.W., Kennedy L.R. Using Classification Tree and Logistic Regression Methods to Diagnose Myocardial Infarction. — <http://groups.csail.mit.edu/medg/people/hamish/medinfo-chris.pdf> (s.a.).