ОЛЕКСІЙ БОРОВСЬКИЙ,
кандидат соціологічних наук, асистент кафедри галузевої соціології Київського національного університету імені Тараса Шевченка

СЕРГІЙ ЛІТВІНОВ,
кандидат соціологічних наук, асистент кафедри галузевої соціології Київського національного університету імені Тараса Шевченка

Специфіка застосування методу дерев рішень в аналізі масиву даних на прикладі порівняльного дослідження

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.

Вступ

Останнім часом в інформаційному просторі України та Росії спостерігається суттєве зростання інтересу до проблем міждержавної взаємодії

двох країн і ролі громадської думки в цих відносинах. Актуалізація цього явища зумовлена як посиленням політичних суперечностей, так і низкою соціальних причин. У книжці, присвяченій дослідженню національно-громадянських ідентичностей і толерантності в Росії й Україні, серед соціальних чинників названо дедалі більші міжетнічні суперечності і соціальну дезорієнтацію людей за умов ціннісно-нормативної невизначеності й аномійної деморалізованості значної частини населення [Национально-гражданские идентичности, 2007: с. 25]. Таке становище більшості населення, безумовно, має впливати на оцінку в масовій свідомості місця і ролі міждержавних відносин. Практичний інтерес для нас становить аналіз громадської думки у двох країнах, а також методологічні особливості аналізу даних порівняльного соціологічного дослідження, реалізованого авторами цієї публікації.

Громадська думка певною мірою детермінована постійними повідомленнями українських і російських соціологічних центрів, котрі, як правило, подають кількісний аналіз одновимірних і двовимірних розподілів відповідей респондентів на запитання. Такі результати, втім, не дають цілісної картини співвідношення між двома об'єктами порівняння. Непоміченою залишається безліч соціальних чинників, першою чергою соціокультурних відмінностей і національних особливостей формування громадської думки.

За таких умов практичного значення набуває метод аналізу, застосовуваний у процесі класифікації великої кількості неоднорідних соціальних даних. У пропонованій статті здійснено спробу обґрунтування евристичного потенціалу **методу "дерев класифікації" (або "дерев рішень")** в аналізі масиву даних порівняльного соціологічного дослідження "Українці і росіяни: погляд один на одного", проведеного у 2008 році за ініціативи Інституту вивчення Росії.

Що таке "дерева класифікації"?

"Дерева класифікації" (classification trees) — порівняно молодий метод data meaning, одна із евристичних процедур глибинного аналізу даних. Перші кроки в цьому напрямі були зроблені наприкінці 50-х років XX століття Говлендом і Гантом. Засадовою стосовно методу дерев вважається пізніша праця Ганта, Меріна і Стоуна "Індуктивні експерименти" (Experiments in Induction), опублікована 1966 року. Уже в 1980-х і особливо у 1990-х алгоритми дерев класифікації стали популярним інструментом біологічних і медичних досліджень, а також мовою моделювання процесу прийняття рішень у науках про управління (див.: [Деревья классификации, s. a.]). Програмні продукти, що реалізують цей метод, у наш час закріпилися в наборі засобів "добування даних".

Іноді *classification trees* відносять до алгоритмів так званого інтелектуального аналізу, що передбачає діалоговий режим і автоматизацію процесу пошуку оптимального рішення (див.: [Classification, s. a.]). У компетенції користувача залишається коректне формулювання завдання, вибір найадекватніших статистичних критеріїв, контроль процесу автоматизованого опрацювання й інтерпретація отриманих результатів.

Метод "дерев класифікації" поєднує переваги алгоритмів, реалізовуваних на сучасній обчислювальній техніці, з творчою участю людини у підготовці вихідних даних, формулюванні гіпотез, у теоретичному осмисленні продукту автоматизованої класифікації — ґрафа ("дерева") рішення. Ця особливість має як плюси, так і мінуси. До перших слід віднести гнучкість методу стосовно вихідних даних, можливість використання різних статистичних критеріїв для класифікації, наочність і добру інтерпретовуваність дерева рішень. До других — статистичну "слабість" результату, відсутність критеріїв надійності класифікації даних, функцію розподілу яких було б вивчено і табульовано. Відтак, метод дерев класифікації слід вважати розвідницьким. Його не можна використовувати у традиційному конфірматорному підході щодо доведення статистичних гіпотез. Навпаки, результати автоматичної класифікації полегшують формулювання їх. Одначе коло завдань цього методу набагато ширше за його суто технологічне застосування. Головне завдання дерева рішення — виявити й візуалізувати приховану категоріальну структуру даних, властиву їм, так би мовити, самим по собі, до аналітичного розчленування статистичним скальпелем. Тому коректне використання *classification trees* дає змогу не лише заощадити масу часу і ресурсів, а й досягти якісно іншого рівня пояснення емпіричних залежностей (див.: [Берестнева, Муратова, 2004]). Пощаблева класифікація об'єктів за багатьма змінними-предикторами, реґресія залежної змінної, формулювання кількісних умов добору об'єкта в одну із заздалегідь виокремлених груп за спостережуваними значеннями тестових змінних — це далеко не повний перелік застосувань методу дерев. Що стосується сутності й різновидів методу, то ми відсилаємо читача до відповідної методологічної літератури (див., напр.: [Эффективная сегментация, s.a.; Classification, s.a.; Tsien, Fraser et al., s.a.]).

Основна ідея методу

Спинімося коротко на головній ідеї дерев рішень. Вона полягає в такому. Нехай задано множину з ознак, квантифікованих числовими змінними — інтервальними, порядковими чи номінальними. Одну із цих ознак (вона має бути категоріальною) ми розглядаємо як залежну, а решту $n - 1$ — як предиктори варіації її значень. Взаємний вплив незалежних змінних одна на одну нас не цікавить. Ми маємо намір виокремити змінну, яка дає змогу щонайкраще згрупувати об'єкти, які різняться за залежною змінною (за цільовим параметром). Іншими словами, знайти змінну, групування за якою вможливлює вирізнення підмножин об'єктів, котрі максимально різнитимуться за варіацією цільового параметра всередині підмножин. Знайшовши таку змінну, ми розглядаємо отримані k_1 підмножин як цільові параметри другого рівня, а решту $n - 2$ незалежних змінних — як предиктори цільових параметрів. Потім процедура повторюється в кожному із k_1 випадків. Ми отримуємо k_2 цільових параметрів третього рівня тощо k_i параметрів $i + 1$-го рівня. Ґраф підмножин розгалужується доти, доки групи виокремлюваних об'єктів стануть занадто малими чи будуть вичерпані всі $n - 1$ вихідних предикторів. При цьому на дереві відображаються лише ті

групи об'єктів (і класифікаційні змінні), котрі значимо різняться за варіацією залежної ознаки. Відповідно, ті "гілки" i-го рівня, які не вдасться розділити на значимо відмінні підмножини жодним із $n - i$ предикторів, уриваються.

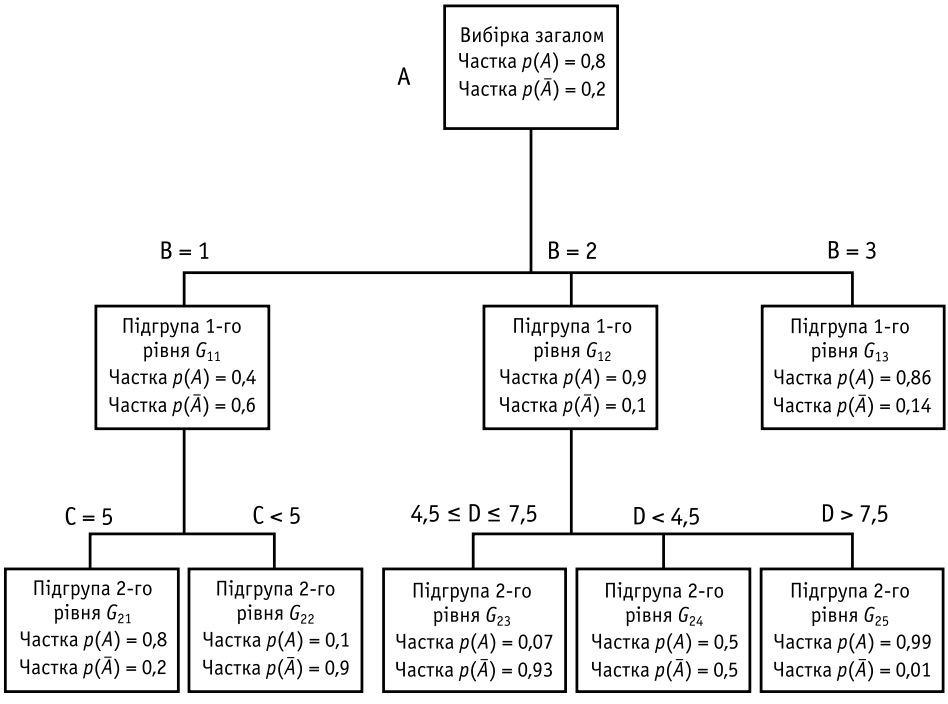


Рис. 1. Умовний приклад дерева рішення, побудованого за методом CHAID

На наведеному ґрафі залежна змінна A — дихотомічна. У теорії дерев рішень вона має назву цільової змінної (параметра) або мітки класу. Цільова змінна є вершиною розгалуження. Виокремлювані алгоритмом підгрупи утворюють вузли ґрафа. Вузли виокремлюються на підставі умови (правила) добору значень незалежної змінної атрибута. Так, у дереві рішення на рис. 1 вузли першого порядку виокремлюються атрибутом B за правилом: якщо $B(j) = 1$, то об'єкт j належить підмножині G_{11}; якщо $B(j) = 2$ — підмножині G_{12}; якщо $B(j) = 3$, об'єкт належить до підмножини G_{13}. Підмножини першого вузла цілковито вичерпують собою вихідну множину (вибірку) $G(A) = G_{11} \cup G_{12} \cup G_{13}$, а правило описує всі значення атрибута B. Номінальна змінна B набуває значення 1, 2, 3; порядкова C — цілі значення від 1 до 5; інтервальна змінна D змінюється в межах від 0 до 10. Підібране алгоритмом правило розподіляє область значень D на три інтервали, що не перетинаються. Відгалуження вузлів першого рівня утворюють вузли другого рівня, тож дочірні підмножини, утворені за специфічним для кожного вузла першого порядку правилом, цілком вичерпують собою вихідну множину. На рисунку 1 гілка уривається, а вузли G_{11} і G_{12} розщеплюються на вузли другого

рівня за правилом для атрибута C (G_{11}) та атрибута D (G_{12}). Кінцеві вузли дерева G_{13}, G_{21}, G_{22}, G_{23}, G_{24}, G_{25} називаються вузлами рішення, або поетичніше — листами. Інтерпретація дерева рішення у цьому умовному прикладі також доволі проста. Найбільшу "поділову силу" стосовно частки ознаки A має атрибут B, що утворює вузли першого порядку. Тому його можна вважати найбільш значимим для варіації A. Інакше кажучи, розподіл A більше залежить від B, ніж від C чи D, точніше, залежить від B першою чергою (про "вплив" тут можна говорити лише в нестрогому й дуже широкому смислі слова). Крім того, за ґрафом легко виокремити різноманітні підгрупи із відмінними середніми частками цільового параметра. Максимум $p(A) = 0{,}99$ можна спостерігати в підгрупі G_{25}, що виділяється за правилом "$B = 2$ і $D > 7{,}5$". Абсолютний мінімум $p(A) = 0{,}99$ досягається в підгрупі (правило "$B = 2$ і $4{,}5 \le D \le 7{,}5$"). Слід також звернути увагу на мінімум $p(A) = 0{,}1$ у листі G_{22} ("$B = 1$ і $C < 5$"). Інші листи можна впорядкувати за міткою класу між мінімумом і максимумом.

Що стосується якості отриманого дерева рішення, то воно має дві складові — точність і надійність. Точність класифікації природним чином можна оцінити за відсотком правильно класифікованих об'єктів. В окремих випадках, наприклад при аналізі медичних даних, важливість правильної класифікації неоднакова для різних вузлів. Для врахування цих відмінностей використовують поняття апріорної ймовірності й ціни помилки класифікації [Деревья классификации, s. a.]). Ми не будемо їх розглядати; зауважимо лише: якщо обрати пропорційні величині класів апріорні ймовірності, а ціну помилки для всіх класів вважати однаковою, то мірою якості класифікації буде частка правильно класифікованих об'єктів. Другу складову якості рішення, надійність, оцінити куди складніше. Статистичних критеріїв для цього просто не існує. У праці [Ростовцев, s. a.] пропонується використовувати бутсреп, методики розмноження вихідної вибірки, щоби на підставі обчислювальних процедур, а не граничних апроксимацій перевірити сталість деревоподібної класифікації, а отже – її надійність.

Сфера застосування, вимоги і можливості "дерева класифікації"

Як використовувати результати аналізу дерева рішення? Основних застосувань три.

1. Опис даних. Отриманий ґраф зручно використовувати замість багатьох таблиць для унаочнення структури даних.

2. Класифікація об'єктів і побудова ієрархії змінних-критеріїв класифікації. Зручність дерева рішення для цієї мети очевидна.

3. Якщо мітка класу континуумальна, дерева рішень дають змогу встановити залежність цільової змінної від незалежних предикторів. До цього класу належать завдання чисельного завбачення значень цільової змінної (регресія).

Можна бачити, що сфера застосування методу "дерев класифікації" перетинається із методами дискримінантного аналізу (якщо цільова змінна дихотомічна), кластерного аналізу, дисперсійного і порядкового регресійного аналізу. Але його перевага крім більшої наочності полягає ще й у можливості одночасного розв'язання кількох задач на підставі одного дерева. Крім того, метод передбачає меншу формалізацію й конкретизацію початкових умов, що робить його гнучкішим і привабливішим для практичного використання. Ті самі переваги забезпечують перспективність "дерев рішень" як інструменту соціологічного аналізу анкетних даних [Ростовцев, s. a. ; Толстова, 2000; Украинское общество, 2007].

Стислий опис даних, як і побудова емпіричної класифікації, належить до найважливіших проблем опрацювання даних, якщо дані являють собою набір множини змінних різного рівня квантифікації, залежності і відношення між ними *a priori* не визначені. Тому на першому етапі опрацювання — до висування статистичних гіпотез — доречна розвідувальна стратегія аналізу. Однією з можливих її реалізацій є застосування групи методів classification tree. Деякі автори рекомендують використовувати "дерева рішень" там, де необхідно отримати однозначні рекомендації на підставі емпірично обчислених правил, наприклад для видачі кредитів, оперативної діагностики хворих тощо [Национально-гражданские идентичности, 2007; Classification, s.a.].

Базові алгоритми

Нині вже запропоновано низку критеріїв, за якими можна оцінити значимість відмінностей, а також алгоритмів побудови ґрафа класифікації. Існує чимало алгоритмів, які реалізують "дерева рішень", наприклад, NewId, ITrule, CN2 тощо. Але найпоширенішими є такі алгоритми (див.: [Деревья классификации, s. a.; Деревья решений, s.a.; Эффективная сегментация, s.a.]):

- **CHAID (CHi-squared Automatic Interaction Detector)**. Розробник — Ґ.В.Кас (1980). "Метод побудови дерев рішень, в якому для отримання оптимальної розбивки використовують критерій зв'язку між категоріальними змінними χ^2 (у разі, якщо цільова змінна є кількісною, використовують F-критерій). Початково цільова змінна і змінні-предиктори можуть бути як кількісними, так і категоріальними, проте кількісні предиктори при побудові дерева перетворюються на категоріальні (кількістю категорій можна управляти)" [Толстова, 2000]. Рідше використовують алгоритми FACT (Loh & Vanichestakul, 1988), THAID (Morgan & Messenger, 1973) або AID (Morgan & Sonquist, 1963).

- **Exhaustive CHAID (Вичерпний CHAID)**. Модифікація методу CHAID. "Його перевагою є те, що в процесі побудови дерева аналізується більша кількість можливих розбивок, а вадою — повільніша швидкість роботи. Цей метод накладає на типи цільової змінної та предикторів такі само обмеження, що й метод CHAID" [Эффективная сегментация, s.a.].

- **C&RT (Classification And Regression Trees),** дослівно — метод класифікації і побудови дерев регресії, запропонований Л.Брейманом та ін. (1984). На відміну від двох описаних вище методів ґрунтується не на статистичних критеріях, а на зменшенні неоднорідності підгруп (вузлів). Для аналізу можна використовувати як кількісні, так і категоріальні цільові змінні і змінні-предиктори. Найліпший результат досягається тоді, коли всі змінні в аналізі є кількісними.

- **QUEST (Quick, Unbiased, Efficient Statistical Trees)**, тобто "швидкі, незміщені, ефективні статистичні дерева" (Loh & Shih, 1997). У цьому методі для вибору предикторів застосовують різноманітні критерії залежно від типу потенційного предиктора. Метод дає змогу уникати зміщень, пов'язаних із вибором предикторів із великою кількістю категорій. Цільова змінна в цьому разі має бути категоріальною. Змінні-предиктори можуть бути як кількісними, так і категоріальними.

- **C4.5.** Розробник — Р. Квінлан (1993). Алгоритм побудови дерева рішень, в якому кількість розгалужень вузла не обмежена. Не призначений до роботи з безперервним цільовим полем, тому розв'язує лише завдання класифікації.

Особливість усіх названих алгоритмів, що визначає специфіку методу дерев рішень, полягає в тому, що коли один раз обрано атрибут, за яким було зроблено розбивку на підмножини, то алгоритм не дає змоги повернутися назад і вибрати інший атрибут, який би давав ліпшу розбивку. Тому на етапі побудови не можна сказати, чи вможливить обраний атрибут оптимальну розбивку.

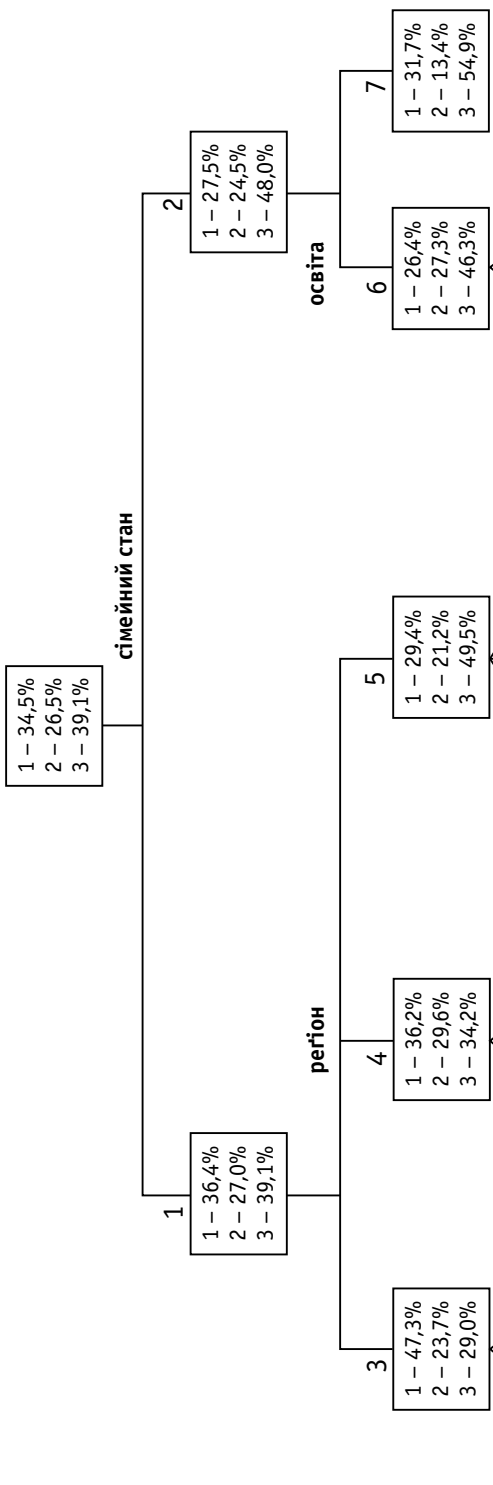
Приклад застосування аналізу дерев до виокремлення критеріїв емпіричної класифікації респондентів

Наш досвід застосування алгоритму Tree Analysys у SPSS 13.0 показує високу ефективність методу дерев класифікації в опрацюванні складних масивів даних соціологічних досліджень. Метод був реалізований нами в перебігу опрацювання даних другої хвилі порівняльного дослідження "Українці і росіяни: погляд один на одного", проведеного на замовлення Інституту вивчення Росії. У Росії опитування проводила компанія "GfK RUS" від 27 червня до 11 липня 2008 року, в Україні — компанія "GfK Ukraine" від 19 червня до 7 липня 2008 року.

Опитування респондентів проводили методом особистого інтерв'ю за місцем проживання. Метою опитування було виявлення найбільш наближених оцінок стану міждержавних відносин між двома країнами. Вибіркова сукупність побудована за схемою багатоступеневої вибірки, отриманої методом випадкового добору (в Росії — 2196 інтерв'ю, в Україні — 1313). Теоретична статистична похибка вибіркового оцінювання частки біноміальної ознаки із розподілом 50% : 50% за довірчої ймовірності $p = 0{,}95$ для України не перевищує 2,7%, для Росії — 2,1%.

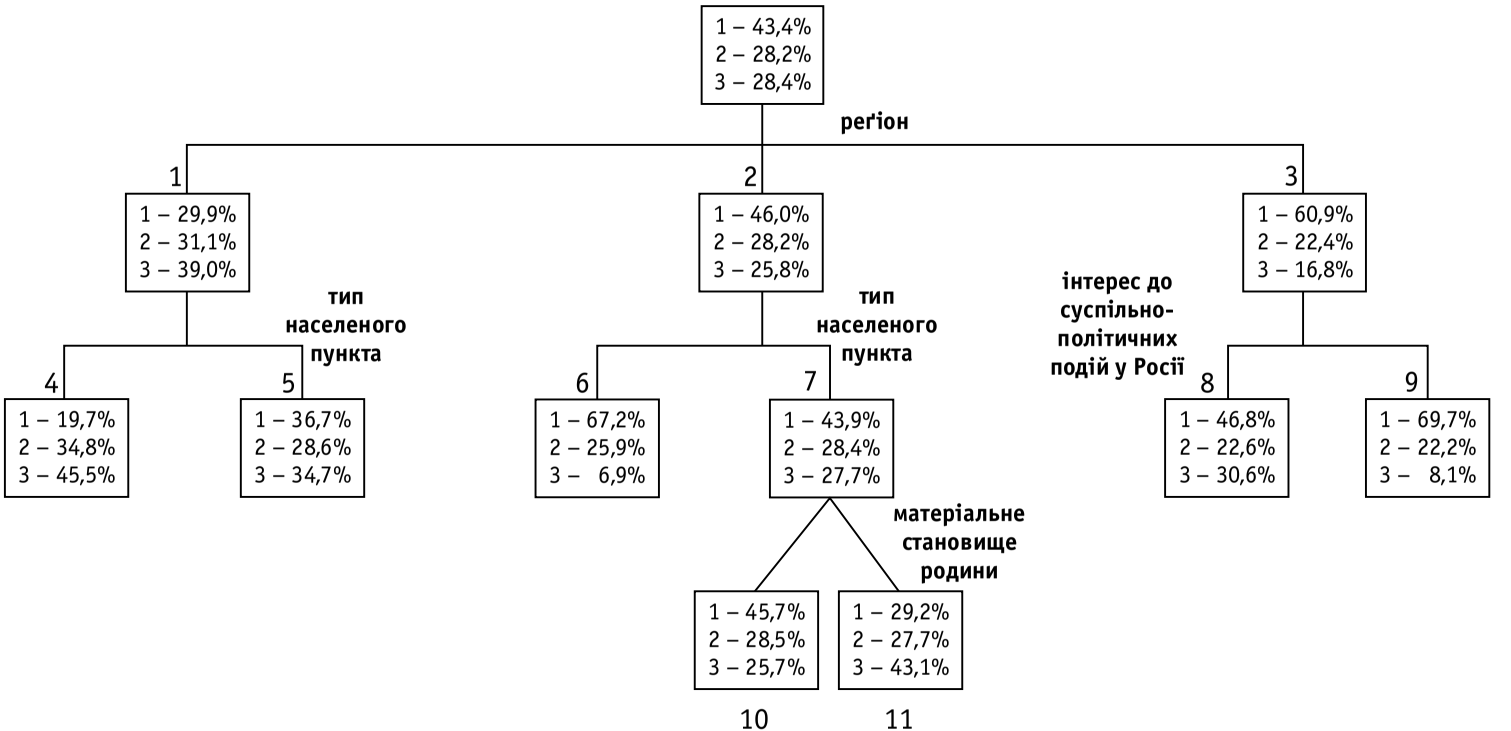
Одним із головних завдань було обчислення інтеґрального індексу добросусідства (ІД). ІД будували на основі ще 6 індексів: трьох простих — базового індексу відносин (БІВ), індексу динаміки відносин між країнами (ІДВК), індексу динаміки відносин між народами (ІДВН) і одного складного — індексу інтересу до іншої країни, її політичного, економічного і культурного життя (ІІІ). При цьому нам було важливо зрозуміти: 1) від яких саме чинників найбільшою мірою залежить ІД і 2) які групи респондентів характеризуються полярними значеннями індексу. Використовувати для цього описову статистику і перевірку гіпотез для побудови класифікації було б неефективно, адже у масиві одних лише соціально-демоґрафічних змінних налічувалося 9. Якщо навіть згрупувати дані, то на базі 9 змінних утворюється не менше $2^9 = 512$ ґрадацій. До того ж переважна частина цих градацій за реалізованих обсягів вибірки була б недостатньо наповненою. Що стосується гіпотез про вплив, то ми не вважали себе достатньо компетентними для вичерпного формулювання їх. Першу частину задачі можна було розв'язати шляхом логістичної реґресії. Але в масиві були як категоріальні, так і кількісні змінні, котрі можна розглядати як потенційні предиктори індексу добросусідства. Крім того, паралельно необхідно було виконати завдання побудови класифікації, виокремлення критеріїв, за якими різняться групи респондентів із високим і з низьким показником ІД. Зазначеним вимогам задовольняла методика CHAID алгоритму дерев рішень у SPSS 13.0. Перш ніж застосувати методику до наших даних, ми побудували залежну змінну, значення якої були обчислені як факторні значення (factor scores). Як індикатори для факторного аналізу ми відібрали змінні індексів БІВ, ІДВК, ІДВН та ІІІ. Виявилося, що найліпшим чином варіацію індикаторів описує двофакторна модель, у якій до першого фактора ввійшли БІВ, ІДІК і ІДВН, а до другого — три змінні-компоненти ІІІ. Оцінювання індексу добросусідства здійснювали на основі першого фактора, що відображає базовий рівень оцінювання респондентами україно-російських відносин. Відповідно до методики обчислення факторних значень шляхом реґресії було розраховано підсумкову безперервну змінну з нормальним розподілом значень від –3 до 3. Перед побудовою дерева класифікації вона була перетворена на категоріальну шляхом розбивки на терцильні інтервали. Значення "1" шкали відповідає нижньому терцилю (погані відносини між державами), "2" — середньому терцилю (нейтральні відносини), а "3" — верхньому терцилю (добрі відносини). Саме цю сконструйовану змінну було взято як залежну в Tree Analysis. Множина залежних змінних містила всі соціально-демоґрафічні ознаки та змінні, на підставі яких ми розраховували індекси ІІІ і ІІК (симетричний ІІІ індекс інтересу до власної країни, див. додаток). Оскільки потенційні предиктори являли собою змінні різних типів і нам була потрібна класифікація, де б усі ґрадації одного предиктора розташовувалися на одному рівні розгалуження дерева рішення, було обрано методику CHAID. Мінімальна наповненість підгруп була визначена у 50 одиниць.

Отримані нами дерева є показовими і за українською, і за російською вибірками. Класифікаційне дерево за російською вибіркою (див. рис. 2) дають змогу правильно класифікувати 71% респондентів, за українською



1 — одружений (заміжня); живуть разом; розлучені; живуть окремо; удівець (удова)
2 — неодружений (незаміжня)
3 — Південний регіон; Далекий Схід
4 — Поволжя; Центральний регіон; Сибір; Урал
5 — Північно-Західний регіон; м. Москва
6 — немає вищої освіти
7 — є вища освіта
8 — середньодушовий дохід сім'ї
9 — тип населеного пункту
10 — тип населеного пункту
11 — тип населеного пункту
12 — до 4000 руб.
13 — 4000 руб. і більше
14 — село, селище; місто із населенням 100–499 тис. жителів
15 — місто із населенням до 100 тис. жителів; місто із населенням понад 1 млн жителів
16 — село, селище; місто із населенням 100 тис. — 1 млн жителів
17 — місто із населенням понад 1 млн жителів
18 — місто із населенням до 100 тис. жителів
19 — міста всіх типів
20 — село, селище

Рис. 2. Класифікаційне дерево з оцінки україно-російських відносин (Росія)



1 — Київ; Північний і Центральний регіони
2 — Західний і Східний регіони
3 — Південний регіон
4 — місто із населенням до 50 тис. жителів; місто із населенням понад 500 тис. жителів
5 — місто із населенням 51–499 тис. жителів
6 — місто із населенням 51–100 тис. жителів
7 — місто із населенням до 50 тис. жителів; місто із населенням понад 100 тис. жителів
8 — мало цікавлюся; важко відповісти; зовсім не цікавлюся
9 — дуже цікавлюся; загалом цікавлюся
10 — погане чи середнє матеріальне становище
11 — відмінне матеріальне становище ("вистачає на все")

Рис. 3. Класифікаційне дерево з оцінки україно-російських відносин (Україна)

(див. рис. 3) — 74%. Ключовою диференціювальною ознакою російських респондентів є сімейний стан, точніше належність до групи неодружених (другий вузол ґрафа, Node 2). Оцінка неодруженими респондентами україно-російських відносин вища, ніж загалом за вибіркою: факторні значення з верхнього терциля зустрічаються серед них на 9% частіше — у 48% респондентів проти 39,1% загалом за масивом. Аналізуючи соціально-демоґрафічні характеристики групи неодружених, можна дійти висновку, що насправді ця ознака маркує вікові відмінності: 82% неодружених становлять люди, молодші за 30 років. Найбільш оптимістичні серед них дві підгрупи: люди із вищою освітою (добрими вважають відносини між Росією й Україною 55%, а поганими — 32%) і жителі села або селища без вищої освіти (60% і 11% відповідно). Більшість респондентів розшаровується за іншими ознаками, реґіональною і поселенською. "Оптимісти" мешкають у Москві й у містах із населенням до 100 тис. Південно-Західного реґіону, "песимісти" — у Південному реґіоні й на Далекому Сході. Відтак, можна зробити висновок, що відносно вище україно-російські відносини оцінюють респонденти, які належать до заможних і соціально оптимістичних груп, а також так звані "прості" люди з низьким показником соціального цинізму.

Якщо звернутися до аналізу української вибірки, то там ситуація дещо інша. По-перше, реґіоналізм в Україні не лише є головним чинником розшарування оцінок відносин між країнами, а й вирізняє кількісно більш диференційовані групи, ніж це можна спостерігати на російській вибірці. Причому виходить парадоксальна річ: різні за всіма соціокультурними параметрами Західний і Східний реґіони опинилися в дереві рішень в одному вузлі (Node 2). Обидва реґіони демонструють "нормально погану" оцінку відносин між країнами. Найбільше "песимістів" у містах із населенням 51–100 тис. жителів, що являють собою соціально-депресивні соціуми ("песимістів" на 60% більше, ніж "оптимістів"). "Оптимісти" локалізовані в селах, селищах і містах із населенням до 50 тис. жителів і в містах із населенням понад 100 тис. До "оптимістів" належать респонденти з високим фінансовим добробутом, які не відчувають ускладнень у задоволенні найважливіших матеріальних потреб. Найбільша різниця в межах України спостерігається між оцінками респондентів із Південного реґіону, з одного боку, і Києва, Північного і Центрального реґіонів — з іншого. У першому випадку число "оптимістів" відноситься до числа "песимістів" як 1 : 3,6, у другому "песимістів" більше за "оптимістів" у 1,35 раза. "Оптимісти" тут — це мешканці малих міст або Києва, які помірковано цікавляться культурно-спортивним життям в Україні. Разом із тим до числа крайніх "песимістів" у Південному реґіоні належать ті, хто виявляє підвищений інтерес до суспільно-політичного життя в Росії. На наш погляд, це дає підстави говорити про соціокультурну детермінацію оцінок українськими респондентами відносин між Україною і Росією. Погляд "песимістів" Півдня України ніби звернений у бік Росії як референтного (ба навіть "свого") політичного простору. "Оптимісти" із Центральної та Південної України, навпаки, звернені у бік власного культурного простору. Можливо, оптимістичне сприйняття відносин між країнами пов'язане саме з аполітичністю представників цієї групи. Таким чином, можна стверджувати, що для України найважливішими де-

термінантами оцінки відносин із Росією є регіональний і соціокультурний чинники, а в Росії більш значимими є соціально-демоґрафічні характеристики респондентів. Реґіоналізм в Україні означає набагато більше, ніж у Росії. Ці висновки, як і емпіричний портрет "оптимістів" і "песимістів", куди складніше було б отримати, застосовуючи традиційні методи аналізу двовимірних розподілів. На нашу думку, евристичний потенціал методу аналізу дерев виявляється саме на етапі узагальнення та групування даних, у перебігу побудови емпіричних типологій. Хоча, безумовно, цей підхід можна використовувати й для того, щоби спробувати "побачити" приховану структуру даних на етапі висування попередніх гіпотез, тобто як допоміжний експлораторний інструмент. Соціологам доведеться ще чимало зробити в плані аналізу ефективності застосування методу класифікаційних дерев і його конкретних алгоритмів щодо даних соціологічних досліджень, а також зі з'ясування оптимальних умов цього цікавого методу.

ДОДАТОК

Змінні-предиктори, застосовувані при побудові класифікаційних дерев для ознаки "Оцінка відносин між Україною і Росією"

1. ЧИ ЦІКАВИТЕСЯ ВИ ПОДІЯМИ, ЩО ВІДБУВАЮТЬСЯ В УКРАЇНІ? (суспільно-політичні події)
2. ЧИ ЦІКАВИТЕСЯ ВИ ПОДІЯМИ, ЩО ВІДБУВАЮТЬСЯ В УКРАЇНІ? (економічні події)
3. ЧИ ЦІКАВИТЕСЯ ВИ ПОДІЯМИ, ЩО ВІДБУВАЮТЬСЯ В УКРАЇНІ? (культурно-спортивне життя)
4. ЧИ ЦІКАВИТЕСЯ ВИ ПОДІЯМИ, ЩО ВІДБУВАЮТЬСЯ В РОСІЇ? (суспільно-політичні події)
5. ЧИ ЦІКАВИТЕСЯ ВИ ПОДІЯМИ, ЩО ВІДБУВАЮТЬСЯ В РОСІЇ? (економічні події)
6. ЧИ ЦІКАВИТЕСЯ ВИ ПОДІЯМИ, ЩО ВІДБУВАЮТЬСЯ В РОСІЇ? (культурно-спортивне життя)
7. СТАТЬ
8. ВІК
9. ЯКИЙ НАВЧАЛЬНИЙ ЗАКЛАД ВИ ЗАКІНЧИЛИ ОСТАННІМ?
10. ЯКЕ ВИСЛОВЛЕННЯ НА ЦІЙ КАРТЦІ НАЙЛІПШЕ ОПИСУЄ ВАС І ВАШУ СІМ'Ю?
11. КИМ ВИ ПРАЦЮЄТЕ ЗАРАЗ?
12. ЯКИМ БУВ ДОХІД ВАШОЇ СІМ'Ї МИНУЛОГО МІСЯЦЯ У РОЗРАХУНКУ НА ОДНОГО ЧЛЕНА СІМ'Ї
13. СІМЕЙНИЙ СТАН У ДАНИЙ ЧАС

14. РЕҐІОН
15. РОЗМІР І ТИП НАСЕЛЕНОГО ПУНКТУ
16. КЛАСИФІКАЦІЯ ESOMAR

Література

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.).