УДК 316:519.2

ДЕРЕВЬЯ КЛАССИФИКАЦИЙ КАК ОДИН ИЗ СПОСОБОВ АНАЛИЗА СОЦИОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

Е. Е. Фомина

Кандидат технических наук, доцент,
ORCID 0000-0002-1028-0750,
e-mail: f-elena2008@yandex.ru,
Тверской государственный технический университет,
г. Тверь, Россия

THE CLASSIFICATIONS TREES AS ONE OF THE METHODS FOR ANALYSIS OF SOCIOLOGICAL INFORMATION

E. E. Fomina

Candidate of Technical Sciences, assistant professor,
ORCID 0000-0002-1028-0750,
e-mail: f-elena2008@yandex.ru,
Tver State Technical University,
Tver, Russia

Abstract. Multidimensional reconnaissance methods allow researchers to solve problems of finding patterns and dependencies in large arrays containing the values of variables describing a certain phenomenon or object. One of the most flexible methods for finding relationships between variables is the classification trees. The article considers the possibilities of the classification tree method in the processing of sociological information, in particular, the results of a survey on the topic «Charity».
Keywords: multidimensional reconnaissance analysis; classification trees; analysis of the survey results.

Методы многомерного разведочного анализа данных стали находить широкое применение в социально-экономических науках. Их основное преимущество заключается в том, что исследователь имеет возможность проверить априорные предположения о структуре зависимостей между переменными, описывающими какое-либо явление или объект и выдвинуть предварительные гипотезы о природе взаимосвязей между ними.

К методам многомерного разведывательного анализа относятся факторный, кластерный, дискриминантный анализ, многомерное шкалирование, анализ соответствий, надежность и позиционный анализ, логлинейный анализ, деревья классификации (ДК) и другие [2, 4, 7, 9–10].

К наиболее гибким методам поиска зависимостей между переменными, описывающими некий объект или явление, относится метод ДК, который используется для прогнозирования принадлежности объектов к тому или иному классу значений зависимой переменной, измеренной в категориальной шкале, на основе значений независимых переменных, которые могут быть как категориальными, так и порядковыми, и интервальными [1, 3].

Цель настоящей статьи – рассмотрение возможностей метода деревьев классификации для анализа социологической информации.

Метод характеризуется построением ДК, состоящего из корневого узла, содержащего всю выборку, дочерних и родительских узлов, а также терминальных узлов, т.е. окончательных узлов, которые далее не делятся. Каждой вершине соответствует правило, согласно которому объекты относятся к тому или иному классу [1, 3, 5–6, 8]. Алгоритм построения дерева классификации включает в себя выбор критерия точности прогноза; выбор метода построения дерева классификации; определение оптимального размера дерева и кросс проверку построенного дерева [1, 3, 5–6, 11].

Рассмотрим применение метода ДК для анализа социологической информации, представляющей собой результаты опроса на тему «Благотворительность». Целью опроса было выяснить, как жители города относятся к благотворительности и насколько активно они принимают участие в тех или иных благотворительных акциях. Опросник, в том числе, включал в себя следующие вопросы:

1. Оказывали ли Вы благотворительную помощь за последние пять лет? Варианты ответа: 0 – да; 1 – нет.

2. Пол респондента: 1 – мужской; 2 – женский.

3. Ваш возраст (количество полных лет).

4. Ваше образование? Варианты ответа: 1 – неполное среднее; 2 – полное среднее; 3 – профессионально-техническое с неполным средним образованием; 4 – профессионально-техническое с полным средним образованием; 5 – среднее специальное образование; 6 – неполное высшее; 7 – высшее.

5. Ваше занятие в настоящее время? Варианты ответа: 1 – работаю; 2 – учусь; 3 – нахожусь на пенсии по выслуге, по возрасту; 4 – нахожусь на пенсии по инвалидности; 5 – веду домашнее хозяйство; 6 – нахожусь в отпуске по беременности, по уходу за ребенком; 7 – безработный, ищу работу; 8 – не работаю и не ищу работу.

6. Ваш доход за последний месяц?

Анкета разработана на кафедре социологии и социальных технологий Тверского государственного технического университета. Выборка составила 1001 респондент, для обработки было отобрано 749 полных опросных листов, содержащих ответы на все вопросы.

Цель анализа – выделить и охарактеризовать группы респондентов, принимающих и не принимающих участие в благотворительности.

В качестве зависимой переменной выступал первый вопрос, в котором респонденты высказывали свое желание или нежелание принимать участие в благотворительных акциях. Остальные вопросы выступали в качестве независимых переменных.

Обработка результатов проводилась в программе STATISTICA [11]. В качестве метода построения дерева использовался метод C&RT. В качестве правила остановки использовалось остановка по отклонению [1].

Результат применения метода – дерево классификации, позволяющее провести наглядную интерпретацию результатов (рис. 1). Дерево содержит 6 терминальных вершин и 5 решающих правил. Прокомментируем его, начиная с корневой вершины, в которой выборка делится на две группы в зависимости от дохода: если доход респондента менее либо равен 33 524 руб., то он попадает в группу респондентов, которые не участвуют в благотворительности (вершина 2, 615 человек), в противном случае – в группу принимающих участие в благотворительности (вершина 3, 134 человека).

Вершина 2 в свою очередь в зависимости от дохода разделяется на две группы: респонденты с доходом менее либо равным 27 470,6 руб., не принимающие участие в благотворительности (вершина 4, 405 человек), и с доходом более 27 470,6 руб., принимающие участие (вершина 5, 209 человек).

Вершина 5 разделяется на две терминальные вершины в зависимости от пола:

для женщин (8 вершина, 112 человек) характерно участие в благотворительности, а для мужчин (9 вершина, 97 человек) – нет.

Вершина 4 в зависимости от дохода разделяется на две группы: респонденты с доходом менее либо равным 25 822,5 руб., не принимают участие в благотворительности (вершина 6, 315 человек), и с доходом более 25 822,5 руб., принимают участие (вершина 7, 91 человек).

Вершина 6 разделяется на две терминальные вершины в зависимости от образования: для людей с высшим образованием характерно участие в благотворительности (10 вершина, 74 человека), в противном случае – нет (11 вершина, 241 человек).

Значимость предикторов распределяется следующим образом: наиболее значимый (ранг 100) – доход, следующий по значимости – пол (ранг 65), далее образование (ранг 49) и возраст (ранг 37), самый наименее значимый предиктор – возраст (ранг 14).

Таким образом, решающим фактором, оказывающим влияние на участие респондентов в благотворительности, является доход. Также можно сделать вывод, что для женщин более характерно участие в благотворительности, чем для мужчин. Фактор образования тоже оказывает влияние на построение решающих правил: люди с невысоким доходом и высшим образованием принимают участие в благотворительности, в отличие от респондентов, не имеющих высшего образования.

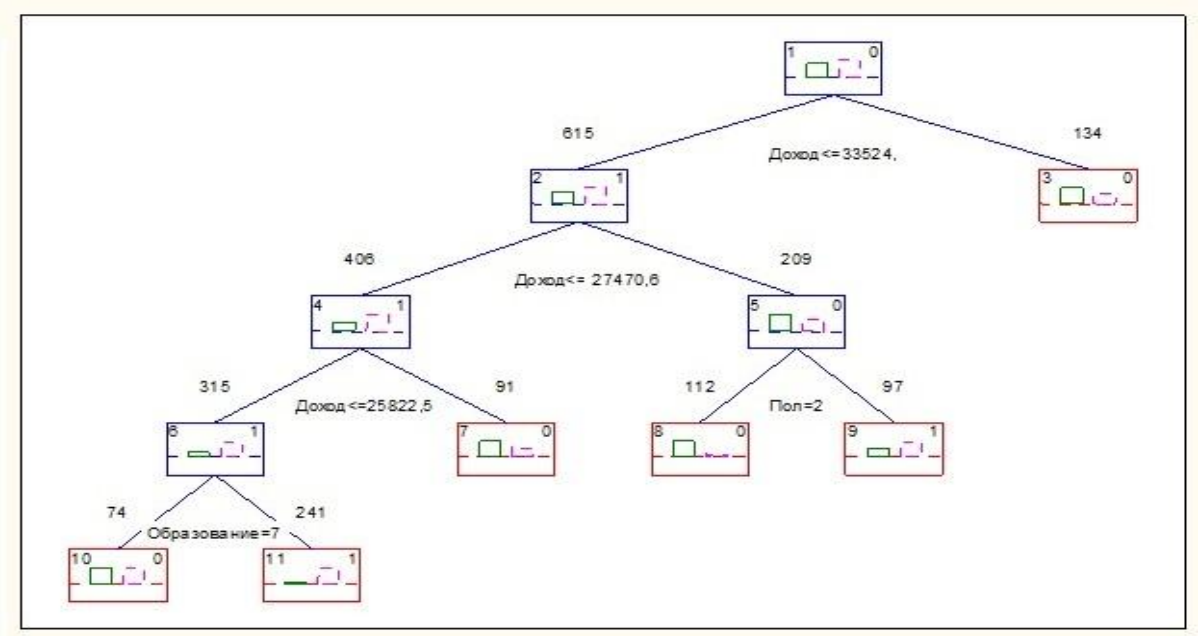

Рис. 1. **Дерево классификации**

Ошибки классификации на обучающей выборке составили 1,8 % для класса 0 и 1,7 % для класса 1. При проведении кросс-проверки доля ошибочно классифицированных респондентов составила 4,5 %, что говорит о хорошем качестве классификации.

В статье рассмотрен алгоритм метода деревьев классификации. Продемонстрированы его возможности при обработке результатов опроса на тему «Благотворительность».

Библиографический список

1. Анализ статистических данных с использованием деревьев решений. Режим доступа: http://math.nsc.ru/AP/datamine/decisiontree.htm (дата обращения 22.04.2018).
2. Аптон Г. Анализ таблиц сопряженности. – М. : Финансы и статистика, 1982. – 143 с.
3. Бова А. Деревья решений как техника добычи данных // Социология: теория, методы, маркетинг. – 2002. – № 1. – С. 128–136.
4. Девятко И. Ф. Методы социологического исследования. Екатеринбург: Издательство Уральского университета, 1998. – 208 с.
5. Деревья классификации. Режим доступа: https://docplayer.ru/20542703-Derevya-klassifikacii.html (дата обращения 18.03.2018).
6. Толстова Ю. Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. – М. : Научный мир, 2000. – 352 с.
7. Трофимов Д. А. Логлинейный анализ таблиц мобильности: обзор основных моделей // Социология. – 2008. – № 26. – С. 119–138.
8. Фомина Е. Е. Использование методов многомерной статистики для анализа социальной и экономической информации // Экономика. Социология. Право. – 2018. – № 2 (10). – С. 61–67.
9. Фомина Е. Е. Факторный анализ и категориальный метод главных компонент: сравнительный анализ и практическое применение для обработки результатов анкетирования // Гуманитарный вестник. – 2017. – № 10 (60). – С. 3.
10. Фомина Е. Е., Жиганов Н. К. Методика обработки результатов анкетирования с использованием методов многомерной и параметрической статистики // Вестник Пермского национального исследовательского политехнического университета. Социально-экономические науки. – 2017. – № 1. – С. 106–115.
11. Электронный учебник по STATISTICA. Режим доступа: http://statistica.ru/textbook/ (дата обращения 20.04.2018).

Bibliograficheskij spisok

1. Analiz statisticheskix danny`x s ispol`zovaniem derev`ev reshenij. Rezhim dostupa: http://math.nsc.ru/AP/datamine/decisiontree.htm (data obrashheniya 22.04.2018).
2. Apton G. Analiz tablicz sopryazhennosti. – M. : Finansy` i stati-stika, 1982. – 143 s.
3. Bova A. Derev`ya reshenij kak texnika doby`chi danny`x // Sociologiya: teoriya, metody`, marketing. – 2002. – № 1. – S. 128–136.
4. Devyatko I. F. Metody` sociologicheskogo issledovaniya. Ekaterinburg: Izdatel`stvo Ural`skogo universiteta, 1998. – 208 s.
5. Derev`ya klassifikacii. Rezhim dostupa: https://docplayer.ru/20542703-Derevya-klassifikacii.html (data obrashheniya 18.03.2018).
6. Tolstova Yu. N. Analiz sociologicheskix danny`x. Metodologiya, deskriptivnaya statistika, izuchenie svyazej mezhdu nominal`ny`mi priznakami. – M. : Nauchny`j mir, 2000. – 352 s.
7. Trofimov D. A. Loglinejny`j analiz tablicz mobil`nosti: obzor osnovny`x modelej // Sociologiya. – 2008. – № 26. – S. 119–138.
8. Fomina E. E. Ispol`zovanie metodov mnogomernoj statistiki dlya analiza social`noj i e`konomicheskoj informacii // E`konomika. Sociologiya. Pravo. – 2018. – № 2 (10). – S. 61–67.
9. Fomina E. E. Faktorny`j analiz i kategorial`ny`j metod glavny`x komponent: sravnitel`ny`j analiz i prakticheskoe primenenie dlya obrabotki rezul`tatov anketirovaniya // Gumanitarny`j vestnik. – 2017. – № 10 (60). – S. 3.
10. Fomina E. E., Zhiganov N. K. Metodika obrabotki rezul`tatov anketirovaniya s ispol`zovaniem metodov mnogomernoj i parametricheskoj statistiki // Vestnik Permskogo nacional`nogo issledovatel`skogo politexnicheskogo universiteta. Social`no-e`konomicheskie nauki. – 2017. – № 1. – S. 106–115.
11. E`lektronny`j uchebnik po STATISTICA. Rezhim dostupa: http://statistica.ru/textbook/ (data obrashheniya 20.04.2018).