ТЕХНИЧЕСКИЕ НАУКИ

УДК 005

Д.И. Лебеденко
студент
С.В. Пантюхов
студент
О.Н. Юркова
к.э.н., доцент кафедры ИТ
ФГБОУ ВО "БГИТУ"
г. Брянск, Россия

СОВЕРШЕНСТВОВАНИЕ МЕТОДОВ ОБРАБОТКИ И АНАЛИЗА ДАННЫХ

Любые методы обработки данных, так или иначе, используются для структурирования и анализа существующей информации. Задач по анализу информации много, однако, в этой статье рассмотрены методы, которые эффективно работают для решения задач по структурированию данных с большим количеством разнородных параметров. Например, для более эффективного продвижения товаров массового потребления на рынок имеет смысл сегментировать потребителей на группы по определенным параметрам: пол, возраст, семейное положение, доход семьи и так далее. Для этого существует набор математических методов, которые позволяют установить закономерности в данных - в случае с анализом потребителей такой закономерностью будут характерные группы потребителей. Часто случается, что формальные методы анализа позволяют получить неожиданные новые знания - например, при исследовании клиентов одной из гостиниц выяснилось, что все клиенты-пенсионеры, проживающие в гостинице, имеют доход свыше 800 долларов.

Типы данных

Данные, которые могут быть использованы для анализа, бывают четырех типов:

1. Численные данные (стоимость товара; 100 рублей).
2. Интервальные данные (доля рынка компании; 5 %).
3. Ранговые данные (лояльность потребителя; напиток Coca-Cola нравится больше, чем Pepsi-Cola).
4. Номинальные данные (профессия потребителя; ученый, военный, врач).

Все данные, которые подходят под один из этих типов, могут быть проанализированы с помощью формальных методов. Любой набор данных может быть адекватно представлен комбинацией перечисленных типов.

Количество данных

Для того чтобы работали большинство методов, желательно иметь более 30 событий (малая выборка). Этого количества событий обычно достаточно для получения информации, что в данной выборке наблюдается статистический эффект. Однако для разделения на группы необходимо иметь уже гораздо большее число событий - примерно 30, умноженное на

число групп. Например, для более-менее правильного разделения потребителей на 3 группы желательно иметь более 100 респондентов. Несомненно, для разных задач и методов количество событий может быть разным, и какую-то информацию можно извлекать уже из 10 событий, однако здесь действует общее правило статистики: чем данных больше, тем лучше.

Методы анализа и обработки данных. Кластерный анализ

Термин "кластерный анализ" (впервые ввел Tryon в 1939 г.) в действительности включает в себя набор различных алгоритмов классификации. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как разбить данные на группы с близкими значениями параметров. Например, при сегментации рынка можно кластеризовать потребителей по двум параметрам - цены и качества. Допустим, компания - производитель автомобилей провела опрос потребителей, в котором задавала два вопроса: "За какую цену Вы готовы купить автомобиль?" и "Оцените качество автомобиля X по 50-балльной шкале" (несколько странный вопрос, однако в качестве иллюстрации он вполне подходит). В результате опроса были получены следующие данные (см. схему; данные в табличной форме не носят информативный характер):

№ участника опроса	Цена, тыс. $	Качество автомобиля X
1	27	19
2	11	46
3	25	15
4	36	27
5	35	25
6	10	43
7	11	44
8	36	24
9	26	14
10	26	14
11	9	45
12	33	23
13	27	16
14	10	47

Кластерные технологии в исследованиях

Если посмотреть на диаграмму (так называемая диаграмма рассеяния) "цена - качество", представленную на рис. 1, то сразу будут видны группы потребителей:

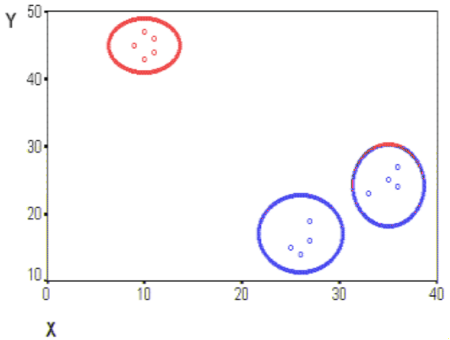

Рис. 1. **Соотношение цена-качество**

Владея этой информацией, каждой группе потребителей можно предложить именно то, что необходимо именно этой группе, и за счет этого увеличить уровень продаж компании. Разумеется, в реальной жизни кластеры, различимые глазом, встречаются нечасто, гораздо чаще бывают ситуации, когда все результирующие параметры смешиваются в одну "кучу" (см. рис.2)

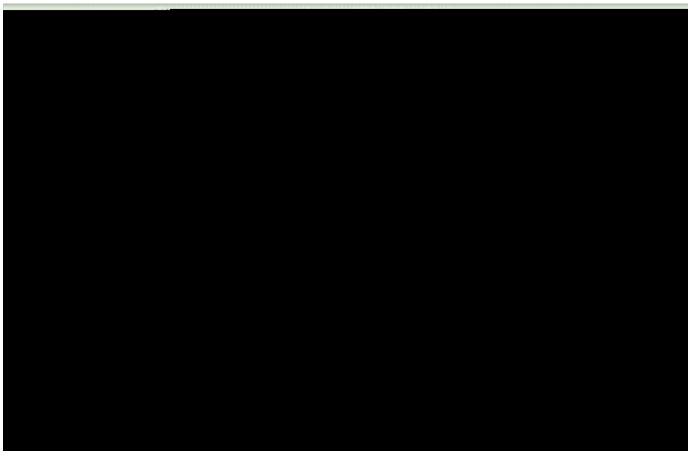

Рис. 2. **Диаграмма рассеяния**

Особенно часто это встречается, когда анализируемых параметров не два, а несколько десятков (кластерный анализ не ограничивает число анализируемых параметров, поэтому можно рассматривать всю проблему комплексно). Глазом кластеры выделить не получится, однако с помощью алгоритмов кластерного анализа это сделать можно. Для проведения кластерного анализа, кроме сбора данных, необходимо определить две вещи: на какое количество кластеров необходимо разделить данные и как определить меру сходства в данных. Например, все предприятия России можно кластеризовать по географическому признаку на 10 кластеров. Тогда мера сходства будет определяться коммуникационной близостью предприятий друг к другу. В более сложных случаях можно применять другие меры сходства, которые подробно описаны в литературе по кластерному анализу. Существует много разных мер сходства, наиболее употребительны из них порядка десяти, но подробно останавливаться на мерах сходства не будем.

Факторный анализ

В случае наличия большого числа параметров (более 100) имеет смысл сгруппировать параметры и анализировать уже не каждый параметр в от- дельности, а группы параметров как единый комплексный параметр (фактор). В основе факторного анализа лежит идея о том, что за сложными взаимосвязями явно заданных признаков стоит относительно более простая структура, отражающая наиболее существенные черты изучаемого явления, а "внешние" признаки являются функциями скрытых общих факторов, определяющих эту структуру. Например, для анализа структуры экономического роста России можно проанализировать все макроэкономические параметры, предварительно объединив их в группы. Одним из таких факторов будет являться ВВП. Объединение параметров можно делать вручную, эмпириче-

Научный и инт

данные в целом. Проверка адекватности обучения. Проводится анализ полученных результатов на данных, которые не входили в обучающую выборку. Осуществляется ручной контроль результатов работы нейронной сети.

Пример задач, которые ставятся перед нейронными сетями

В работе А. Горбаня приводится следующий пример использования нейронных сетей: нейронная сеть обучалась предсказывать результаты выборов президента США по ряду экономических и политических показателей. Обученные сети были минимизированы по числу входных параметров и связей. Оказалось, что для надежного предсказания исхода выборов в США достаточно знать ответы всего на пять вопросов, приведенных ниже в порядке значимости:

1. Была ли серьезная конкуренция при выдвижении от правящей партии? Кластерные технологии в исследованиях
2. Отмечались ли во время правления существенные социальные вол- нения?
3. Был ли год выборов временем спада или депрессии?
4. Произвел ли правящий президент значительные изменения в поли- тике?
5. Была ли в год выборов активна третья партия?

От использования остальных признаков нейронная сеть отказалась. Более того, эти пять "симптомов" политической ситуации в стране входят в распознающее правило двумя "синдромами". Пусть ответы на вопросы кодируются числами: +1 - "да" и 1 - "нет". Первый синдром есть сумма ответов на вопросы 1, 2, 5. Его естественно назвать синдромом политической нестабильности (конкуренция в своей партии, плюс социальные волнения, плюс дополнительная оппозиция). Чем он больше, тем хуже для правящей партии. Второй синдром - разность ответов на вопросы 4 и 3 (политическое новаторство минус экономическая депрессия). Его наличие означает, что политическое новаторство может, в принципе, уравновесить в глазах избирателей экономический спад. Результаты выборов определяются соотношением двух чисел - значений синдромов. Простая, но достаточно убедительная политологическая теория, чем-то напоминающая концепцию то ли Макиавелли, то ли Ленина ("единство партии прежде всего, оно является важнейшим слагаемым политической стабильности"). Этот пример показывает несколько достаточно важных вещей с точки зрения использования нейронных сетей:

1. Нейронные сети могут выделять значимые факторы.
2. Факторы могут быть сгруппированы в "синдром" (см. факторный анализ).
3. Исследование осмысленности работы нейронной сети остается за исследователем.

Деревья решений

Первые идеи создания деревьев решений восходят к работам Ховленда (Hoveland) и Ханта (Hunt) конца 50-х годов XX века. Однако основополагающей работой, давшей импульс для развития этого направления, явилась книга Ханта (Hunt E.B.), Мэрина (Marin J.) и Стоуна (Stone P.J.) "Experiments in Induction", увидевшая свет в 1966 г. Деревья решений - это способ представления правил в иерархической последовательной логической структуре, который позволяет соотнести объект или ситуацию на входе с одним или несколькими выходными (терминальными) узлами. Под правилом понимается логическая конструкция, представленная в виде "если... то". Рассмотрим следующую задачу: необходимо построить решающее правило по выдаче кредита физическим лицам. В этом случае дерево решений может выглядеть следующим образом (рис. 3):

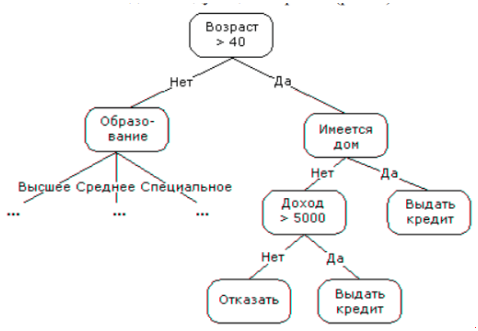

Рис. 3. **Четкое дерево решений**

В предыдущем примере приведено так называемое четкое дерево решений. Не меньший интерес представляют вероятностные деревья решений (рис. 4), в которых каждый параметр принятия решения может входить в результирующее решение с некоторой вероятностью:

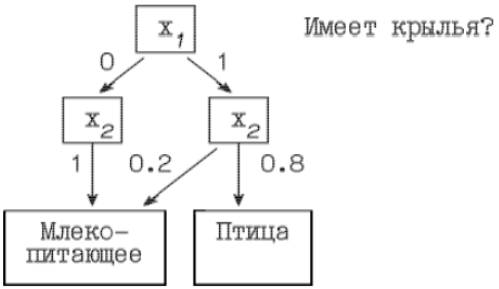

Рис. 4. **Вероятностное дерево решений**

Метод деревьев решений может помочь при принятии сложного решения, на которое влияют десятки параметров. Деревья решений широко применяются во многих областях деятельности:

1. Банковское дело. Оценка кредитоспособности клиентов банка при выдаче кредитов.
2. Промышленность. Контроль за качеством продукции (выявление дефектов), испытания без разрушений (например, проверка качества сварки) и т. д.
3. Медицина. Диагностика различных заболеваний.
4. Молекулярная биология. Анализ строения аминокислот.
5. Консалтинг. Компания McKinsey использует деревья решений (issue tree, термин McKinsey) для консультаций своих клиентов.

Это далеко не полный список областей, где можно использовать деревья решений. Не исследованы еще многие потенциальные области применения этого инструмента.

Регрессионный анализ

Основной целью регрессионного анализа является определение наличия и характера связи между переменными (в простейшем случае строится зависимость y(x) исходя из пример-

ной формы кривой). Несколько лет назад американский Институт стратегического планирования провел исследование "Маркетинговая стратегия и уровень прибыли", в котором рассматривалось влияние наиболее значимых переменных на уровень прибыли компании. Выяснилось, что график зависимости рентабельности от доли рынка выглядит следующим образом (рис. 5):

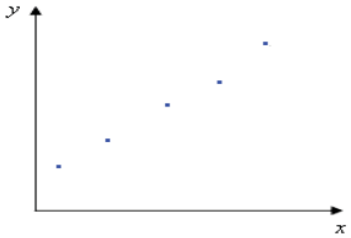

Рис. 5. **График зависимости рентабельности от доли рынка**

Невооруженным взглядом видно, что это прямая, однако точные ее параметры помогает установить регрессионный анализ. Регрессионный анализ широко используется в офисном пакете Excel, который предоставляет возможность исследовать не только линейные, но и другие, более сложные зависимости (в Excel это называется построением линий трендов).

Дискриминантный анализ

Дискриминантный анализ - это инструмент статистики, который используется для принятия решения о том, какие переменные разделяют возникающие наборы данных. Например, некий исследователь в области образования решает исследовать, какие переменные относят выпускника средней школы к одной из трех категорий: (1) поступающий в колледж, (2) поступающий в профессиональную школу или (3) отказывающийся от дальнейшего образования или профессиональной подготовки

между двумя различными выборками (например, при сравнении пар близнецов), и если эта связь существует, то сопровождается ли увеличение одного показателя возрастанием (положительная корреляция) или уменьшением (отрицательная корреляция) другого [1-6].

Заключение

Задачи восстановления зависимостей активно изучаются уже более 200 лет, с момента разработки К. Гауссом в 1794 г. метода наименьших квадратов. В математической статистике с этого времени было разработано огромное количество методов и инструментов анализа данных. В данной работе описаны методы, которые наиболее широко используются во всем мире и во всех областях прикладной науки для анализа данных - физике, биологии, экономике, психологии и др. В настоящее время компьютеры играют большую роль в математической статистике. Они используются как для расчетов, так и для имитационного моделирования, в частности в методах размножения выборок и при изучении пригодности асимптотических результатов.

Список литературы

1. Ефремова М.В. Сегментация потребителей гостиничных услуг // Маркетинг в России и за рубежом. 2002. № 2.

2. Толстова Ю.Н. Измерение в социологии: Курс лекций. М.: ИНФРА-М, 1998.

3. Гаскаров Д.В., Шаповалов В.И. Малая выборка. М.: УРСС, 1978 // [Электронный ресурс] http://ru.wikipedia.org/wiki/Выборка

4. Дюран Б., Одедд П. Кластерный анализ / Пер. с англ. М.: УРСС, 1977.

5. Садохина Е.Ю. Факторный анализ структуры экономического роста России и Беларуси за 1991-2002 гг. // Научные труды ИНП РАН.

6. Сергей Колесников. Зачем нужны нейронные сети? // [Электронный ресурс]КОМПЬЮТЕР-ИНФОРМ, http://www.ci.ru/inform15_05/ p_08.htm

УДК 621

Т.И. Петров
ассистент кафедры "Электроснабжение промышленных предприятий"
Казанский Государственный Энергетический Университет
Р.Р. Сахапов
Инженер
АО "Средне-Волжский Транснефтепродукт"
г. Казань, Россия

ОПТИМАЛЬНЫЙ СРОК СЛУЖБЫ ЭЛЕКТРООБОРУДОВАНИЯ СИСТЕМ ЭЛЕКТРОСНАБЖЕНИЯ

Оптимальный срок службы электрооборудования зависит от всех этапов жизненного цикла изделия. В данной статье будет рассмотрен только один из этапов - эксплуатация. Важным элементом эксплуатации является ремонт и понятие "срок проведения мониторинга". Для определения оптимального срока мониторинга возможно использование дополнительного параметра - экономические риски.

При планировании мониторинга определяется сроки его проведения. Основанием для планирования сроков служит нормативная периодичность, которая определяется, как зафиксирова