МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ
ГОРНЫЙ УНИВЕРСИТЕТ

СЕРИЯ «ПРАКТИЧЕСКАЯ СТАТИСТИКА ДЛЯ ГОРНЫХ ИНЖЕНЕРОВ»

Л.Х. Гитис

СТАТИСТИЧЕСКАЯ КЛАССИФИКАЦИЯ И КЛАСТЕРНЫЙ АНАЛИЗ

МОСКВА

ИЗДАТЕЛЬСТВО МОСКОВСКОГО ГОСУДАРСТВЕННОГО ГОРНОГО УНИВЕРСИТЕТА

2003

УДК 519.251:622
ББК 22.172
Г 46

Гитис Л.Х.
Г 46 Статистическая классификация и кластерный анализ. — М.: Издательство Московского государственного горного университета, 2003. — 157 с.: ил.
ISBN 5-7418-0010-6

Посвящена теории распознавания образов и одному из методов ее реализации — кластерному анализу. В сжатом виде представлены основные идеи кластерного анализа и показаны сферы его приложения в горных, экономических, социологических и других исследованиях. Описанные методы кластеризации могут быть использованы в реальных задачах. В алгоритмах достаточно подробно рассмотрена вычислительная часть.

Несмотря на то, что кластерный анализ является эффективным и удобным инструментом классификации, а также весьма распространен в практических исследованиях, публикаций на эту тему на русском языке очень мало, а существующие малоинформативны. Предлагаемая Вашему вниманию книга освещает некоторые основополагающие вопросы кластерного анализа.

Для научных сотрудников, диссертантов и специалистов, работающих в области многомерного статистического анализа.

Табл. 3, ил. 30, список лит. — 29 назв., глоссарий.

УДК 519.251:622
ББК 22.172

ISBN 5-7418-0010-6

СТАТИСТИЧЕСКИЕ МЕТОДЫ В УПРАВЛЕНИИ. КЛАССИФИКАЦИЯ И КЛАСТЕРНЫЙ АНАЛИЗ

Рыночная экономика, несмотря на внешнюю простоту, требует высокой концентрации знаний, интеллекта, умения предвидеть последствия управленческих действий, инициативы и других качеств специалистов, принимающих участие в управлении.

Казалось бы знания экономических и математических теорий здесь необязательны, а сложные статистические методы распознавания образов вообще ни к чему. Но практика свидетельствует, что менеджер, владеющий сложными теориями, более эффективен, чем необразованный управленец.

Для того, чтобы построить систематизированную схему использования различных методов в процессе принятия решений, сформулируем основные содержательные задачи, которые естественным образом возникают в практике управления предприятием, территориальной общностью, отраслью, торгово-коммерческой деятельностью. Поскольку предметом нашего изучения будут статистические теории, им посвятим основное внимание.

Чем же отличается эффективный менеджер от простого управленца? В основном объемом, детализацией и уровнем синтетической обработки используемой информации: если второй тип руководителя использует только личный опыт, да и тот не в полном объеме, то эффективный менеджер перед принятием решения анализирует сходные ситуации, рассчитывает возможные последствия, ищет оптимальные выходы из сложившихся ситуаций. И здесь нельзя обойтись без знания статистических методов. Широко распространено мнение о том, что опытному управленцу для принятия эффективных решений вполне достаточно личного опыта или информации, полученной вербальным путем (на совещаниях, в беседах и т. п.). В простейших условиях это утверждение может быть и справедливым. Но сложная конкурентная среда требует иных знаний и подходов, здесь необходимо умение работать с обобщенной эмпирической, аналитической и синтезированной информацией об управляемых объектах, их аналогах и т. д.

Таким образом эффективность управления производственными, социальными и экономическими процессами в широком масштабе во многом зависит от умения пользоваться всем спектром аналитических и статистических методов. На рис. 1 представлена схема последовательного использования числовых и нечисловых аналитико-статистических методов изучения разнообразной многопараметрической информации. Корректное их использование позволяет глубже понимать сущность происходящих процессов, а также вести осознанный поиск оптимальных решений локальных и условно глобальных задач.

И все-таки, несмотря на разнообразие поставленных задач и методов их решения, основная нагрузка приходится на статистические методы изучения информации. Как видно из приведенной схемы, статистические методы применимы для анализа экономического состояния предприятия, его конкурентоспособности на рынке, социально-производственной привлекательности территориальной общности и производства, а также для решения множества других задач, стоящих перед управленцами всех направлений и рангов.

Для управления немаловажное значение имеет, в каком виде будет представлена информация: необработанном и хаотичном или в классифицированном ранжированном и структурированном. Причем речь идет не только об информации, полученной в результате наблюдений, но и о таких данных, которые являются паспортными для исходных объектов. Объектами изучения могут быть предприятия и их подразделения, производственные мощности и инфраструктура, люди – работники предприятий и члены их семей. И в каждом случае они будут характеризоваться набором показателей: обобщенных и детальных, описывающих сходные явления и специфические черты объектов.

Очевидно, неклассифицированная и неструктурированная информация для управления бесполезна, да и обилие характеристик (многомерность пространства размещения объектов) мешает анализу. Поэтому на первом этапе изучения информацию приходится классифицировать и структурировать. Наиболее удобной формой классификации информации является представление ее в виде неких однородных образов, которые объединяют объекты по принципу близости характеристик. Создание достаточно полной системы образов исходного массива информации позволяет дифференцированно подходить к управлению. Кроме того, однородность образов дает возможность прогнозировать их реакцию (обратную связь) на управляющие воздействия.

Рис. 1. Укрупненная схема анализа и синтеза информационных массивов, используемых для принятия коммерческих решений

Специалисты, работающие в сфере изучения закономерностей и управления минерально-сырьевым комплексом, решают задачи анализа конъюнктуры внутреннего и мирового рынков полезных ископаемых, которые, в свою очередь, подразделяются на задачи ранжирования, оценивания, поиска связей и т. д. Далее возникают задачи многокритериального сравнительного анализа добычных и обогатительных горных предприятий, предлагаемого на рынок продукта, его качества и цены, транспортных технологий и т. д.

Еще один блок задач математической статистики возникает в результате потребности в изучении внутрифирменных связей. Горно-промышленное предприятие расчленяется на десятки основных производств и инфраструктурных подразделений, которые являются объектами управления. Между этими объектами существуют корреляционные зависимости, которые кроме количественных характеристик (коэффициенты корреляции) описываются уравнениями регрессии. Все это позволяет интерполировать и экстраполировать имеющиеся в результате вычислений статистические ряды.

При составлении планов ведения горных работ также используются разнообразные статистические методы. Для подсчета и анализа запасов полезных ископаемых, обобщения разнообразных природных ресурсов применяются методы интерполяции, построения дискриминантных линий, статистические подсчеты содержания массы полезных ископаемых в геометрических блоках горных пород. На основе полученных данных рассчитываются показатели качества рудой массы, решаются задачи усреднения, формируются временные ряды, а затем разрабатываются календарные планы добычи, усреднения и обогащения. В этих задачах важную роль играют методы распознавания незрительных образов.

По какому плану ни проводилось бы исследование информации, первым этапом должны стать ее классификация, фильтрация и представление в упорядоченном и компактном виде. Для решения этих задач и были разработаны методы кластерного анализа. Этап классификации во многом определяет результаты всего исследования. От того, насколько удачно разделены объекты, удалось ли на первом этапе отфильтровать недостоверную информацию, не потеряна ли объективность при группировке детализированных характеристик, зависят точность принимаемых решений и управление, образно говоря, с открытыми глазами.

В то же время нельзя слепо полагаться только на результативность формальных процедур кластерного анализа. Статистические методы классификации требуют осмысления каждого логического этапа вычислений. Для этого на ключевых этапах бывает полезно интерпретировать результаты и к их оценке привлекать экспертов, специализирующихся на содержании исследования, вникающих в суть выполняемых преобразований.

Для понимания общих принципов формулировки и решения широкого класса задач классификации методами кластерного анализа обозначим укрупненно основные стадии обработки информации (рис. 2).

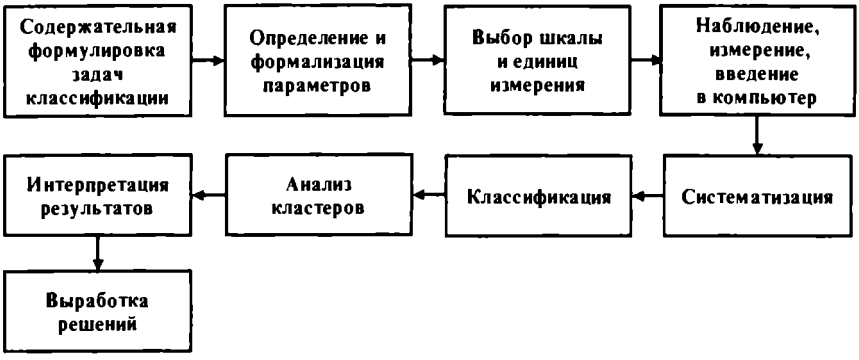

Рис. 2

Глава 1

КЛАССИФИКАЦИЯ, ТЕОРИЯ РАСПОЗНАВАНИЯ НЕЗРИТЕЛЬНЫХ ОБРАЗОВ И ЕЕ РЕАЛИЗАЦИЯ МЕТОДОМ КЛАСТЕРНОГО АНАЛИЗА

1.1. Классификация объектов как необходимое условие деятельности человека

1.2. Логическая модель распознавания незрительных образов, основанная на принципах обучения живых организмов условным рефлексам

1.3. Объективная классификация

1.4. Гипотеза о компактности образов

Глава 1
КЛАССИФИКАЦИЯ, ТЕОРИЯ РАСПОЗНАВАНИЯ НЕЗРИТЕЛЬНЫХ ОБРАЗОВ И ЕЕ РЕАЛИЗАЦИЯ МЕТОДАМИ КЛАСТЕРНОГО АНАЛИЗА

1.1. КЛАССИФИКАЦИЯ ОБЪЕКТОВ КАК НЕОБХОДИМОЕ УСЛОВИЕ ДЕЯТЕЛЬНОСТИ ЧЕЛОВЕКА

Несомненно, классификация – основополагающий процесс в интеллектуальной деятельности человека. Встречаясь с новым явлением, мы стараемся найти ему аналог в известной нам области. Рассматривая группу каких-либо объектов, мы непроизвольно разделяем их на подгруппы близких друг другу элементов. Классификация присутствует при упорядочении известных нам фактов, явлений, предметов.

Конечно же, эти факты и явления должны быть упорядочены прежде, чем появится возможность разобраться в них и понять механизмы управления. Тем более в науке классификация играет весьма значительную роль: тому примерами могут служить теории Менделеева, Линнея, Дарвина. Историки не могли бы без классификации объяснить генезис явлений, происхождение фактических событий, существующий порядок. На основании сказанного можно заключить, что классификация – это фундаментальное понятие науки и практики.

Любопытно, но и в мире животных задачи классификации решаются непрерывно. Хищники классифицируют объекты охоты, стадные – своих и чужих, абсолютно все животные каким-то образом распределяют съедобное и несъедобное. Очевидно, что те живые организмы, которые были неспособны определить группы раздражителей, упорядочить их и найти адекватную реакцию, оказались нежизнеспособными и вымерли.

Отметим, что классификация одних явлений и объектов оказывается инстинктивной деятельностью, а других – полученной в результате обучения. Таким образом, предмет нашего рассмотрения – методы классификации (распознавания образов) не надуманны, а являются вполне естественной областью повседневной и повсеместной деятельности человека при систематизации и оценке явлений и предметов.

Наибольший практический интерес для рассмотрения задач классификации представляют многомерные статистические исследования. Именно многомерность объектов делает их приближенными к реальным проблемам экономики, технологий, социологическим и биологическим задачам. При этом существует очевидная закономерность: сложность объекта и глубина анализа прямо пропорциональны размерности информационного поля.

Среди возможных методов классификации несомненный практический интерес вызывают методы распознавания образов. Здесь мы будем рассматривать методы незрительного распознавания образов, которые изучают числовые и нечисловые переменные и постоянные величины, используют методы математической, статистической и логической их обработки. Одной из ведущих теорий в области распознавания образов является кластерный анализ, благодаря которому решение задач классификации было осуществлено несложными компьютерными методами, а также были получены легко интерпретируемые результаты.

Одной из основополагающих задач классификации является формализация отличия одного объекта от другого. Мы прекрасно знаем, чем отличается автомобиль от трактора, фото от живописи, мужчина от женщины. Но формализовать это знание бывает не очень просто. И если умением компьютера различать тексты, геометрические образы или рисунки сегодня никого не удивить, то найти совпадения или различия нечетко обрисованных объектов, сложноструктурированных множеств или «узнать» неявное течение процесса непросто.

В практике встречаются два основных типа классификации. Простейший случай включает в себя заранее классифицированное пространство (млекопитающие, множество натуральных чисел, элементарные частицы) с известными характеристиками, определяющими принадлежность классу. В этом случае любой новый объект (предмет) можно по совпадению характеристик или причислить к какому-либо классу, или нет. Такая классификация называется «распознаванием с учителем».

Более сложная ситуация возникает при необходимости объективной классификации множества объектов без предварительных подсказок о числе классов, наиболее существенных характеристиках и принципах разделения. Такая классификация называется «распознаванием без учителя» и является основной задачей кластерного анализа.

При классификации сложных объектов возникает проблема выбора наиболее значимых характеристик. Здесь возможны крайние, но малопригодные для реальной деятельности ситуации. Если учитывать все характеристики объектов исходного множества, да еще и очень точно их измерять, то в этом случае каждый объект, скорее всего, будет составлять отдельный класс. Теоретически подобное решение задачи классификации вполне объяснимо, но совершенно непригодно для практики. С другой стороны, возможно такое обобщение характеристик объектов, что они полностью попадут в один общий класс. И этот вариант классификации малопригоден для реального управления по причине примитивной тривиальности.

Таким образом, возникает еще одна фундаментальная задача классификации: выбор оптимального набора характеристик объектов, отвечающий содержательным потребностям управления, систематизации, прогнозирования. Эта задача решается методами факторного анализа, но практическая сторона этих методов значительно сложнее кластерного анализа и, в определенных случаях, более субъективна.

В истории целенаправленной деятельности человека классификация осуществлялась методами, тесно связанными с предметом классификации. Интуитивно-эвристические подходы к классификации были доступны только выдающимся ученым (Менделеев, Линней, Дарвин, Дьюи, Бредфорд) и были результатом их озарения. И основной сложностью при этом был выбор наиболее важной характеристики или нескольких характеристик, по которым и определялось сходство или различие объектов.

И только применение математико-статистических методов позволило находить типологические меры сходства и различия в автоматическом режиме, а иногда даже выполнять эту работу без понимания содержательных основ причин выбора именно тех характеристик, которые бывают названы в результате расчетов.

1.2. ЛОГИЧЕСКАЯ МОДЕЛЬ РАСПОЗНАВАНИЯ НЕЗРИТЕЛЬНЫХ ОБРАЗОВ, ОСНОВАННАЯ НА ПРИНЦИПАХ ОБУЧЕНИЯ ЖИВЫХ ОРГАНИЗМОВ УСЛОВНЫМ РЕФЛЕКСАМ

Присущая каждому животному приобретенная способность классифицировать известные ему объекты может послужить аналогом модели «распознавания с учителем». За основу подобного распознавания принимается простейшая модель обу-

чения живых организмов условным рефлексам. При этом не рассматривается вариант, когда новый объект сравнивается с эталонными для каждого кластера, и в результате выбирается нужный. Такой алгоритм не содержит элементов самообучения или обучения с учителем, что делает его примитивно тривиальным.

Для практики управления и прогнозирования представляет интерес модель обучения распознавания образов. Предположим, что модель не располагает предварительными сведениями о тех объектах, которые нужно классифицировать, или о возможных классах объектов. В таком случае общие характеристики могут быть выведены из сравнения реальных показателей объектов и их синтеза. Процесс обучения, сходный с алгоритмом образования условных рефлексов, заключается в том, чтобы предлагаемая модель адекватно реагировала на новый (неопознанный) объект, узнавала его и идентифицировала с каким-либо классом. На рис. 3 представлена принципиальная модель классификации, построенная по аналогии с выработкой условных рефлексов.

Рис. 3

Если моделью не предусматривается предварительное введение данных о тех объектах, которые далее следует классифицировать, то научиться этому можно только в процессе самой классификации. При этом после классификации начальных объектов модель должна «узнавать» все последующие, одновременно уточняя границы классов. Когда надежность «узнавания» достигнет заранее заданной величины, можно утверждать, что модель пригодна для дальнейшего применения в практических задачах.

Добавим, что получившиеся в результате разбиения классы вполне корректно можно называть некоторыми образами явлений, группой экономических или социальных объектов, набором ошибок каких-либо задач, диагнозов болезней и множества других объединений. Эти образы имеют такие характеристики, которые выражают наиболее существенные черты классифицируемых объектов, но пренебрегают второстепенными, несущественными деталями, способными только нивелировать рассматриваемые объекты.

Именно поиск таких (такого) объективных, в рамках решаемой задачи, свойств, присущих всем рассматриваемым объектам и отсутствующих у других объектов, не входящих в класс, и является наиболее сложной задачей статистической классификации. Модель, вооруженная знанием главных характеристик системы образов, условно может быть названа моделью, у которой выработан устойчивый условный рефлекс «узнавания» своих объектов, причем и в дальнейшем эта модель будет учитывать удачные решения.

1.3. ОБЪЕКТИВНАЯ КЛАССИФИКАЦИЯ

Классификация «без учителя» является весьма эффективным инструментом в начальной стадии изучения разнородных объектов. Действительно, впервые встретив неизученное множество объектов, исследователь задается вопросом: по какому принципу возможна классификация? Что объединяет различные элементы множества, а что позволяет их различать? Конечно, можно прибегнуть к помощи экспертных оценок, которые помогут разделить исходное множество (классификация «с учителем»). Но это разделение будет субъективным и, возможно, ошибочным.

Если же мы хотим разделить множество на группы без предвзятости, без начальных установок и целевых функций, то