Библиографические ссылки

1. Медведев А. В. Теория непараметрических систем. Моделирование // Вестник СибГАУ. 2010. Вып. 4. С. 4–9.

2. Ермаков С. М., Михайлов Г. А. Курс статистического моделирования. М. : Наука, 1976.

3. Фадеева Л. Н. Математика для экономистов: Теория вероятностей и математическая статистика. М. : Эксмо, 2006.

4. Вероятность и математическая статистика : энциклопедия / под ред. Ю. В. Прохорова М. : Большая Рос. энцикл., 2003.

5. Первушин В. Ф., Сергеева Н. А. Генератор случайных чисел, распределенных по логнормальному закону // Решетневские чтения : материалы XIII Междунар. науч. конф. : в 2 ч. Ч. 2 / Сиб. гос. аэрокосмич. ун-т. Красноярск, 2009. С. 448–449.

6. Сергеева Н. А., Стрельников А. В. Генератор случайных чисел, распределенных по закону Вейбулла // Решетневские чтения : материалы XIII Междунар. науч. конф. : в 2 ч. Ч. 2 / Сиб. гос. аэрокосмич. ун-т. Красноярск, 2009. С. 451–453.

V. F. Pervushin, N. A. Sergeeva, A. V. Strelnikov

THE RANDOM SELECTION PRECISION GENERATOR

The algorithm of random value selection with adjusted distributions generation and numerical characteristics is considered (mathematical estimation, dispersion etc.).The algorithm proceedings of high precisions is showed on calculated examples of selection modeling. The selection generation common concept allows to extract this approach to the special category of random value generation algorithms. It was joined to the term "The Random selection precision generator".

Keywords: random value, retribution law, numeric characteristics, event probability, event frequency, selection size, bar chart.

УДК 519.6

В. Б. Бериков

АЛГОРИТМ АДАПТИВНОГО ПЛАНИРОВАНИЯ АНСАМБЛЯ ТАКСОНОМИЧЕСКИХ ДЕРЕВЬЕВ РЕШЕНИЙ*

Рассматривается подход к решению задач кластерного анализа, основанный на применении ансамбля таксономических деревьев решений. Предлагается алгоритм адаптивного планирования ансамбля, использующий расстояния между логическими высказываниями, описывающими кластеры. Приводятся результаты статистического моделирования, подтверждающие эффективность алгоритма.

Ключевые слова: кластерный анализ, ансамбль, дерево решений.

Одной из актуальных проблем кластерного анализа (таксономии, автоматической классификации «без учителя») является группировка объектов, описываемых разнотипными (количественными или качественными) переменными. Другая актуальная проблема связана с повышением устойчивости группировочных решений, так как в большинстве алгоритмов кластерного анализа результаты могут сильно меняться в зависимости от выбора начальных условий, порядка объектов, параметров работы алгоритма и т. п.

Наиболее перспективный подход к кластерному анализу при наличии разнотипных переменных основан на применении деревьев решений, которые позволяют получать легко интерпретируемую логическую модель группировки, выделять наиболее информативные факторы и не требуют задания метрики в разнотипном пространстве. Особенностью логико-вероятностного подхода, основанного на деревьях решений, является возможность не только разбивать заданное множество объектов на кластеры, но и строить иерархическое дерево, описывающее структуру разбиения.

Повысить устойчивость кластеризации можно с помощью ансамблей алгоритмов. При этом используются результаты группировки, полученные различными алгоритмами или одним алгоритмом, но с различными параметрами настройки, по различным подсистемам переменных и т. д. После построения ансамбля проводится нахождение итогового коллективного решения. Такой способ описан, например, в работе [1].

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проекты 08-07-00136а, 10-01-00113а, 09-07-12087-ofi_m) и Междисциплинарного интеграционного проекта Сибирского отделения Российской академии наук № 83.

В работе [2] предложен алгоритм построения коллектива таксономических деревьев решений. При построении ансамбля использовалось введенное в [3] понятие расстояния между логическими высказываниями, описывающими кластеры. В данной статье описано усовершенствование этого алгоритма путем применения методики адаптивного планирования, при которой ведется целенаправленная селекция наиболее перспективных элементов ансамбля.

Основные понятия. Пусть имеется множество $s = \{o^{(1)}, ..., o^{(N)}\}$ некоторых объектов, выбранных из генеральной совокупности. Каждый объект описывается с помощью набора переменных $X_1, \ldots, X_n$. Этот набор может включать в себя переменные разных типов (количественные и качественные, под которыми будем понимать номинальные и булевы, а также порядковые переменные). Пусть D_j обозначает множество значений переменной X_j – некоторый интервал числовой оси в случае количественной переменной или конечный набор значений (имен) в случае качественной переменной. Пусть $D = D_1 \times ... \times D_n$. Обозначим через $x = x(o) = (x_1(o), ..., x_n(o))$ набор наблюдений переменных для объекта *o*, где $x_j(o)$ – значение переменной X_j для данного объекта.

В задаче кластерного анализа требуется разбить объекты на некоторое число кластеров K ($K << N$) так, чтобы заданный критерий качества группировки принял бы оптимальное значение. Под критерием качества обычно понимается некоторый функционал, зависящий от разброса объектов внутри группы и расстояний между группами. Число классов может быть выбрано заранее или не задано (в последнем случае оптимальное количество кластеров должно быть определено автоматически).

Введем понятие таксономического дерева решений. Рассмотрим дерево, в котором каждой внутренней вершине (узлу) соответствует некоторая переменная X_j, а ветвям, выходящим из данной вершины, – определенное высказывание вида $X_j(o) \in E_j^{(i)}$, где *o* – некоторый объект; $i = 1, ..., v$, $v \ge 2$ – число ветвей, выходящих из данной вершины; $E_j^{(1)}, ..., E_j^{(v)}$ – попарно непересекающиеся подмножества множества D_j (интервалы значений в случае количественной переменной). Каждому *k*-му листу (концевой вершине) дерева соответствует группа объектов выборки $C^{(k)} = \{o^{(i_1)}, ..., o^{(i_{n^{(k)}})}\}$, где $n^{(k)}$ – число объектов, входящих в группу, $k = 1, ..., K$. Объекты данной группы удовлетворяют цепочке высказываний, проверяемых по пути из корневой вершины в этот лист, т. е. логическому утверждению вида «Если $X_{j_1}(o) \in E_{j_1}^{(i_1)}$ И $X_{j_2}(o) \in E_{j_2}^{(i_2)}$ И ... И $X_{j_{q_k}}(o) \in E_{j_{q_k}}^{(i_{q_k})}$, то объект *o* относится к *k*-му классу», где q_k – длина данной цепочки, $o \in C^{(k)}$. Под таксоном $T^{(k)}$, соответствующим *k*-му листу дерева, будем понимать прямоугольную подобласть, заключающую объекты из $C^{(k)}$ в многомерном пространстве:

$$T^{(k)} = T(C^{(k)}) = T_1^{(k)} \times ... \times T_j^{(k)} \times ... \times T_n^{(k)},$$

где $T_j^{(k)} = \{X_j(o) \mid o \in C^{(k)}\}$ для качественной переменной X_j; $T_j^{(k)} = [\min\limits_{o \in C^{(k)}} X_j(o); \max\limits_{o \in C^{(k)}} X_j(o)]$ для количественной переменной X_j.

В работе [2] описан регуляризующий критерий качества группировки, основанный на минимизации функционала, зависящего от суммарного относительного объема таксонов и от их числа:

$$Q = \sum_{k=1}^{K} \prod_{j=1}^{n} \frac{|T_j^{(k)}|}{|D_j|} + \alpha \frac{K}{N},$$

где $\alpha > 0$ – регуляризующий параметр, а также предложен алгоритм направленного поиска оптимального дерева, в котором применяется рекурсивная схема перебора различных вариантов.

Построение ансамбля деревьев решений. Пусть с помощью алгоритма построения таксономического дерева решений строится разбиение множества объектов *s* на *K* подмножеств. Для построения каждого варианта случайным образом формируется своя подсистема переменных (либо случайным образом выбираются параметры алгоритма). Под группировочным решением будем понимать набор $G = \{C^{(1)}, ..., C^{(k)}, ..., C^{(K)}\}$, где $C^{(k)} = \{o^{(i_1)}, ..., o^{(i_{N_k})}\}$; N_k – число объектов, входящих в *k*-й кластер, $k = 1, ..., K$. Пусть построен набор группировочных решений $\mathbf{G} = \{G^{(1)}, G^{(2)}, ..., G^{(L)}\}$, где $G^{(l)}$ – *i*-й вариант группировки. Обозначим через $S^{(l)}$ бинарную матрицу размерности $N \times N$, которая водится для *l*-й группировки следующим образом: $S^{(l)}(q, m) = 0$, если объекты $o^{(q)}$ и $o^{(m)}$ принадлежат одному кластеру, $S^{(l)}(q, m) = 1$, если иначе, где *q*, *m* = 1, 2, …, *N*; *l* = 1, 2, …, *L*. После построения *L* группировочных решений формируется согласованная матрица различий

$$S(q, m) = \frac{1}{L} \sum_{l=1}^{L} S^{(l)}(q, m),$$

где *q*, *m* = 1, 2, …, *N*. Величина $S(q, m)$ равна частоте классификации объектов $o^{(q)}$ и $o^{(m)}$ на разные группы в наборе группировок **G**. Близкое к единице значение этой величины означает, что данные объекты имеют большой шанс попадания в одну и ту же группу, а близкое к нулю говорит о том, что шанс оказаться в одной группе у этих объектов незначителен.

После вычисления согласованной матрицы подобия для нахождения итогового варианта группировки можно применять алгоритм построения дендрограммы, который в качестве входной информации использует расстояния между объектами. В предлагаемом нами алгоритме используется аналогичный принцип,

однако вместо бинарной матрицы $S^{(l)}$, в которой отражаются события типа вхождения (либо невхождения) пары объектов в одну и ту же группу, берется более информативная матрица расстояний между кластерами, к которым отнесены объекты. Для вычисления расстояний между кластерами в разнотипном пространстве переменных будем использовать введенное в работе [3] расстояние между логическими высказываниями экспертов.

Пусть в одном и том же варианте группировки имеются два таксона $T^{(s)}$ и $T^{(r)}$, $s, r = 1, \ldots, K$, описываемые некоторыми логическими высказываниями. Введем расстояние между данными таксонами как $\rho(T^{(s)}, T^{(r)}) = f(\rho_1(T_1^{(s)}, T_1^{(r)}), ..., \rho_j(T_j^{(s)}, T_j^{(r)}), ..., \rho_n(T_n^{(s)}, T_n^{(r)}))$, где f – некоторая функция; $\rho_j(T_j^{(s)}, T_j^{(r)})$ – расстояние между j-ми компонентами множеств $T^{(s)}$ и $T^{(r)}$.

Зададим способ вычисления величины ρ_j в зависимости от типа переменной. Если X_j – номинальная переменная, то расстояние определяется как взвешенная мера симметрической разности:

$$\rho_j(T_j^{(s)}, T_j^{(r)}) = \frac{|T_j^{(s)} \Delta T_j^{(r)}|}{|D_j|}.$$

Если X_j – количественная переменная, а интервалы

$$T_j^{(s)} = [a^{(s)}, b^{(s)}], \quad T_j^{(r)} = [a^{(r)}, b^{(r)}],$$

то

$$\rho_j(T_j^{(s)}, T_j^{(r)}) = \frac{|a^{(s)} - a^{(r)}| + |b^{(s)} - b^{(r)}|}{2|D_j|}.$$

В работе [3] было доказано, что для введенного расстояния выполняются все свойства меры, в том числе неравенство треугольника.

В качестве функции f предлагается рассматривать функцию $f_1 = \left(\sum_{j=1}^{n} w_j \rho_j \right)^{1/2}$, где $0 \le w_j \le 1$ – некоторые веса ($\sum_{j=1}^{n} w_j = 1$), либо функцию $f_2 = \max_j \rho_j$, удобную тем, что она не требует задания весов. Для данных функций также выполняются свойства меры.

Таким образом, согласованная матрица различий с учетом расстояний между кластерами определяется как

$$\tilde{S}(q,m) = \frac{1}{L} \sum_{l=1}^{L} \tilde{S}^{(l)}(q,m),$$

где $\tilde{S}^{(l)}(j,m) = \rho(T^{(s)}, T^{(r)})$, $T^{(s)}, T^{(r)}$ – таксоны, к которым принадлежат объекты $o^{(q)}$ и $o^{(m)}$ соответственно.

В алгоритме построения ансамбля [2] выбор подсистемы переменных осуществляется случайно. Однако этот выбор (испытание) можно организовать так, чтобы лучшие по критерию качества группировки варианты имели бы большую вероятность отбора. В основе предлагаемой модификации лежит идея адаптивного случайного поиска наиболее информативной подсистемы переменных, предложенная Г. С. Лбовым.

Первоначально все переменные имеют одинаковый вес, т. е. вероятность отбора в подсистему, используемую при построении дерева. Предполагается, что число переменных достаточно велико. После проведения очередного испытания вес отобранной в корне дерева переменной уменьшается, за счет чего достигается более высокая степень разнообразия ансамбля. Если при этом критерий качества ухудшается, то эта переменная наказывается, т. е. вероятность ее отбора еще более уменьшается пропорционально критерию. В случае когда критерий качества улучшается, соответствующая переменная поощряется, т. е. ее вес уменьшается не столь сильно. То, что выбирается переменная, соответствующая именно корню дерева, объясняется тем, что она является наиболее информативной и оказывает определяющее влияние на дальнейшее ветвление.

Алгоритм планирования состоит из следующих основных шагов.

Шаг 1. Присвоить всем переменным одинаковые веса $P_{X_j} = 1/n, \quad j = 1, ..., n.$

Шаг 2. Выбрать случайным образом d переменных в соответствии с назначенными весами, где d – заданный параметр.

Шаг 3. Построить дерево решений в случайном подпространстве размерности d, в котором X^* – переменная, соответствующая корню дерева. Определить для построенного дерева критерий качества группировки Q.

Шаг 4. Присвоить переменной X^* новый вес:

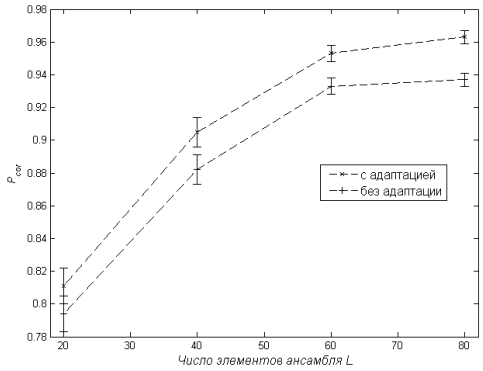
$$P_{X^*} = \max\{P_{X^*} - \tau Q, 0\},$$

где τ – параметр адаптации.

5. Повторить шаги 2…4 до тех пор, пока не будет построено заданное число элементов ансамбля.

Экспериментальное исследование алгоритма. Для определения качества алгоритма была разработана процедура статистического моделирования. Эта процедура состоит в многократном генерировании случайных выборок в соответствии с заданным распределением для каждого класса; построении с помощью исследуемого алгоритма согласованного группировочного решения для каждой выборки; нахождении усредненного по всем выборкам показателя качества. Распределение для каждого из $K = 2$ классов является многомерным нормальным с одной и той же ковариационной матрицей Σ. Число переменных $n = 100$, объем выборки для каждого класса – 25. Вектор математических ожиданий для каждого класса выбирается случайно из множества, соответствующего вершинам единичного гиперкуба. Ковариационная матрица является диагональной: $\Sigma = \sigma \mathbf{I}$, где $\sigma = 1$. Из общего числа переменных 20 являются шумовыми (для них $\sigma = 10$). Номера шумовых переменных выбираются случайно. Каждое дерево строится в слу-

чайно выбранном подпространстве переменных размерности 3. Параметры $\alpha = 2,\ \tau = 0,05$, в качестве функции *f*, задающей вид расстояния между высказываниями, выбрана f_1.



Алгоритм построения коллектива таксономических деревьев решений с адаптацией и без адаптации (указаны 95%-е доверительные интервалы для вероятности правильной классификации)

Качество группировки определяется как частота правильной классификации P_{cor}. Усреднение проводится по 100 случайным выборкам, являющихся реализациями смеси указанных распределений. Оценивается 95%-й доверительный интервал для вероятности правильной классификации. Полученные результаты моделирования позволяют сделать вывод о том, что при достаточно большом числе элементов ансамбля адаптивный алгоритм дает значимо более высокое качество классификации по сравнению с разработанным ранее вариантом, в котором адаптация не используется (см. рисунок).

Таким образом, предложен алгоритм адаптивного планирования ансамбля таксономических деревьев решений, использующий расстояния между логическими высказываниями, описывающими кластеры. Результаты статистического моделирования подтвердили эффективность разработанной процедуры. В дальнейшем планируется применить данный алгоритм в задаче анализа биомедицинской информации, относящейся к устойчивости паразитарной системы клещевого энцефалита, а также для обработки спутниковых и натурных данных.

Библиографические ссылки

1. Strehl A., Ghosh J. Clustering ensembles – a knowledge reuse framework for combining multiple partitions // J. of Machine Learning Research. 2002. Vol. 3. P. 583–617.

2. Бериков В. Б. Кластерный анализ с использованием коллектива деревьев решений // Науч. вестн. Новосиб. гос. техн. ун-та. 2009. № 3 (36). С. 67–76.

3. Лбов Г. С., Бериков В. Б. Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации. Новосибирск : Изд-во Ин-та математики, 2005.

V. B. Berikov

ALGORITHM FOR ADAPTIVE PLANNING OF AN ENSEMBLE OF TAXONOMIC DECISIONS TREES

We suggest an approach to cluster analysis based on the ensemble of taxonomic decisions trees. The adaptive algorithm for the ensemble planning that uses distances between logic statements describing clusters is offered. The results of statistical modeling confirm the efficiency of the algorithm.

Keywords: cluster analysis, ensemble, decisions tree.

УДК 681.3

А. В. Бобров, Е. А. Перепелкин

ВОССТАНОВЛЕНИЕ ИЗОБРАЖЕНИЙ НА ОСНОВЕ СПЕКТРАЛЬНОЙ МАТРИЧНОЙ НОРМЫ

Рассматривается проблема восстановления изображения. Описывается нелинейный гауссовский фильтр, построенный на основе спектральной матричной нормы. Приводятся результаты численных экспериментов, подтверждающие преимущество данного фильтра по сравнению с линейным гауссовским фильтром.

Ключевые слова: восстановление изображения, гауссовский фильтр, спектральная норма матрицы.

Пусть изображение задано в виде неотрицательной вещественной матрицы *A* размером $n \times m$ с элементами $a_{ij} \in [0 \quad a_{\max}]$. Будем считать, что изображение содержит искаженные пиксели. Координаты искаженных пикселей известны, соответствующие элементы матрицы изображения равны нулю. Необходимо восстановить искаженные пиксели.

Для восстановления изображения применяют пространственные и частотные фильтры [1; 2].