УДК 62-50

КОМПЛЕКСНЫЙ АЛГОРИТМ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ И ЕГО ИСПОЛЬЗОВАНИЕ В ЗАДАЧАХ АНАЛИЗА И ПРИНЯТИЯ РЕШЕНИЙ[1]

© Дорофеюк Ю.А.

Институт проблем управления РАН

Abstract. The automatic classification (cluster analysis) complex algorithm, that was especially developed for intellectual complex-organized data handling and decision support problems, is described.

It includes: the m-local optimization algorithm with the predetermined classification performance criterion, the informative parameters selection algorithm, the initial fragmentation algorithm, the missing observation filling algorithm.

Введение

Многие крупномасштабные системы управления, в первую очередь организационно-административные, функционируют в условиях большой информационной размытости и неопределённости. Именно поэтому в последнее время для исследования таких систем стали широко использоваться структурно-классификационные методы, базирующиеся на алгоритмах классификационного анализа данных [1].

В настоящей работе рассматриваются задачи анализа функционирования крупномасштабных систем управления, при этом считается, что такая система состоит из достаточно большого числа объектов, каждый из которых характеризуется многочисленным набором разнородных параметров. Основная идея предлагаемого метода решения этой задачи состоит в том, что исследуются не точные значения параметров, описывающих состояние каждого объекта, а лишь структура взаиморасположения этих объектов в пространстве параметров [2]. Такое интегральное описание объектов, входящих в крупномасштабную систему, позволяет существенно повысить эффективность анализа поведения системы, а также устойчивость и робастность процедур принятия управленческих решений. Для формализации такой задачи используется методология классификационного анализа данных [1].

1. Комплексный алгоритм автоматической классификации

Пусть исследуемая система состоит из n объектов, каждый из которых характеризуется набором из k параметров. Вводится в рассмотрение k-мерное пространство параметров X, в котором каждый объект представляется точкой $x_j = (x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(k)})$, $j = 1, \ldots, n$. Предполагается, что вектор значений параметров x_j достаточно полно характеризует состояние j-го объекта, а это, в свою

[1]Работа выполнена при частичной финансовой поддержке РФФИ, проект 08-07-00349-а.

очередь, означает, что взаиморасположение множества точек $x_1, \ldots, x_n$ в пространстве X отражает реальную структуру (типологию) исследуемого множества объектов. Для выявления такой структуры в работе используется комплексный алгоритм автоматической классификации, специально разработанный для решения таких задач. Комплексный алгоритм включает алгоритмы: m-локальной оптимизации заданного критерия J, выбора информативных параметров, выбора начального разбиения, выбора числа классов, заполнения пропущенных наблюдений. Рассмотрим каждый из этих алгоритмов в отдельности.

1.1. **Алгоритм m – локальной оптимизации.** Вначале опишем работу алгоритма 1-локальной оптимизации. Для простоты изложения рассматривается случай двух классов $r = 2$. Пусть задано начальное разбиение R_0 всех точек классифицируемой выборки $x_1, \ldots, x_n$. Обозначим через $x_j \in A_1$ точки, относящиеся к первому классу, а через $x_j \in A_2$ – ко второму. Алгоритм итерационный, – на каждом шаге рассматривается одна точка из последовательности $x_1, \ldots, x_n, x_1, \ldots, x_n, x_1, .$ («зацикленная» исходная последовательность). Отнесение точки к одному из двух классов обозначается с помощью индекса $\rho(x_j) = \begin{cases} 1, \text{ если } x_j \in A_1 \\ -1, \text{ если } x_j \in A_2 \end{cases}$. Тогда алгоритм 1-локальной оптимизации определяется следующим образом: $\rho\,(x_j) = sign\,[J\,(x_j \in A_1) - J\,(x_j \in A_2)]$.

В итоге точка x_j относится к тому классу, при отнесении к которому, значение критерия J будет больше (если эти значения равны, то для определённости точка относится к классу с меньшим номером). Алгоритм заканчивает работу, если на некотором цикле среди точек $x_1, \ldots, x_n$ не будет сделано ни одной «переброски» точки из класса в класс.

Алгоритм m-локальной оптимизации – это поэтапное применение к выборке алгоритмов s-локальной оптимизации, $s = 1 \div m$. На s-ом этапе алгоритм работает по той же схеме, только на каждом его шаге происходит пробная «переброска» из класса в класс не одной, а s точек. Подсчитывается значение критерия J до и после «переброски», Принадлежность каждой из sточек к классу либо остаётся неименной (J до «переброски» больше, чем после), либо меняется на другой класс – в противном случае. В данном случае цикл – это число шагов, равное числу всевозможных различных наборов, в каждый из которых входит s точек, выбранных из n точек исходной выборки. Доказана сходимость алгоритма за конечное число шагов к локальному максимуму критерия J.

Разработан эвристический алгоритм сокращённого перебора, который на каждом шаге для пробной «переброски» использует s точек в определённом смысле ближайших к границе между классами.

При моделировании и в приложениях в качестве критерия J использовался функционал J_1 средней близости точек в классах, определяемый через потенциальную функцию [3] близости точек x и y:

$$K(x, y) = 1/\{1 + \alpha\, R^p(x, y)\}, \qquad (1)$$

где α и p – настраиваемые параметры алгоритма. Средняя близость точек в классе определяется как:

$$K(A_i, A_i) = \frac{2}{n_i(n_i-1)} \sum_{i=1}^{n_i} \sum_{j>i} K(x_i, x_j), \tag{2}$$

где $K(x_i, x_j)$ определяется формулой (1), n_i– число точек в классе A_i. Тогда критерий J_1 определяется как:

$$J_1 = \sum_{i=1}^{r} \frac{n_i}{n} K(A_i, A_i). \tag{3}$$

1.1.1. *Алгоритм m-локальной оптимизации, одномерный случай.* Необходимо специально отметить частный случай алгоритма m-локальной оптимизации для k=1 (одномерный случай). Дело в том, что одномерный случай имеет уникальное свойство, существенно упрощающее процедуру целенаправленного перебора, используемые при автоматической классификации, а именно: ввиду одномерной упорядоченности классов границей между двумя классами (в детерминированном случае) служит только одна точка, и таких границ может быть не более двух (для крайне правого и крайне левого классов – только одна). Далее описана работа детерминированного (в отличие от общего – размытого) варианта этого алгоритма [4].

Пусть задано начальное разбиение R_0 всех точек классифицируемой выборки $x_1, . \ . \ ., \ x_n$ на r классов. Очевидно, что ввиду упорядоченности классов на оси единственного параметра, на каждом конкретном шаге алгоритма достаточно рассматривать только пару соседних классов, для определённости будем обозначать через A_1 левый из этой пары классов, а через A_2 – правый. Алгоритм содержит mциклов, на s-м цикле (s= 1, ..., m) производится локальная оптимизация классификации, полученной на предыдущем цикле, за счёт процедуры «переброски» s точек из одного класса в другой для каждой пары соседних классов.

На первом цикле производится «переброска» по одной точке. Здесь классификация, полученная на предыдущем цикле, – это начальная классификация R_0. Поясним эту процедуру для первого этапа этого цикла, когда рассматривается пара классов, расположенная в самой левой части диапазона значений x_j. Обозначим через A_1 и A_2 соответственно первый и второй классы начального разбиения R_0. В классе A_1 находится точка $x_j^{1,1,1}$(индексы сверху – номера цикла, этапа и номера класса соответственно), ближайшая к границе рассматриваемой пары классов. Обозначим через $\rho_0\left(x_j^{1,1,1}\right)$ индекс этой точки (для аналогичной точки на s-м цикле это обозначение будет иметь вид $\rho_{s-1}\left(x_j^{s,1,1}\right)$). По построению $\rho_0\left(x_j^{1,1,1}\right)$=1. Затем «перебросим» эту точку в класс A_2 и подсчитаем её индекс на первом цикле:

$$\rho_1\left(x_j^{1,1,1}\right) = sign\left[J\left(x_j^{1,1,1} \in A_1\right) - J\left(x_j^{1,1,1} \in A_2\right)\right], \tag{4}$$

где $J\left(x_j^{1,1,1} \in A_1\right)$ – значение критерия качества классификации J, подсчитанное только для точек классов A_1 и A_2 при условии, что точка $x_j^{1,1,1}$ принадлежит классу A_1, аналогично определяется $J\left(x_j^{1,1,1} \in A_2\right)$. Из (4) следует, что точка $x_j^{1,1,1}$ остаётся в первом классе ($\rho_1\left(x_j^{1,1,1}\right) = \rho_0\left(x_j^{1,1,1}\right) = 1$), если $J\left(x_j^{1,1,1} \in A_1\right) \geq J\left(x_j^{1,1,1} \in A_2\right)$,

и переходит во второй класс ($\rho_1\left(x_j^{1,1,1}\right) = -1$) в противном случае. Если точка $x_j^{1,1,1}$ перешла во второй класс, то аналогичная процедура проделывается с точкой $x_{j-1}^{1,1,1}$, которая является ближайшей к новой границе между классами A_1 и A_2 среди всех точек первого класса (в данном случае – это предыдущая точка классифицируемой последовательности). И так продолжается до тех пор, пока точка $x_{j-l}^{1,1,1}$ не останется в первом классе, т.е. на первом этапе первого цикла из первого класса во второй будут «переброшены» l ближайших к границе точек. Если точка $x_j^{1,1,1}$ осталась в первом классе, то аналогичная процедура проводится с точками второго класса начиная с точки $x_j^{1,1,2}$, которая является ближайшей к границе рассматриваемой пары классов. После того как закончится «переброска» точек из второго класса в первый (если это будет иметь место) либо не произойдёт «переброски» точки $x_j^{1,1,2}$, происходит переход на второй этап первого цикла.

На втором этапе вся последовательность процедур первого этапа повторяется, только через A_1 обозначаются точки, входящие во второй класс после завершения первого этапа первого цикла, а через A_2 – третий класс начального разбиения R_0. И так далее, до тех пор, пока не будут пройдены все (r -1) этапов первого цикла.

На всех этапах s-го цикла описанные процедуры повторяются с точностью до числа «перебрасываемых» точек – «перебрасывается» не по одной, а по s точек, ближайших к границе текущей пары классов. Процедура не применяется для классов A_i, число точек n_i в которых меньше, чем $(s+2)$.

Значение m (глубина перебора) должно выбираться из условия: в классификации, полученной после $(m-1)$-го цикла, должен быть хотя бы один класс, число точек в котором не меньше $(m+2)$. Этим правилом можно воспользоваться для автоматического выбора максимально возможной глубины перебора.

Завершение m-го цикла является окончанием первой итерации. На второй итерации повторяются все процедуры первой, только на первом цикле вместо начального разбиения R_0 используется результирующая классификация первой итерации.

Алгоритм прекращает работу, если в пределах одной итерации не произойдёт ни одной «переброски» точек из класса в класс.

1.2. **Алгоритм выбора информативных параметров.** Этот алгоритм базируется на одном из алгоритмов экстремальной группировки параметров, а именно на алгоритме «квадрат» [3]. В результате его применения получают разбиение исходных k параметров на небольшое (заданное) число групп, а также значения факторов для этих групп. В приложениях используются либо новые интегральные параметры – факторы групп, либо набор параметров, каждый из которых является ближайшим к фактору в соответствующей группе.

В большинстве приложений исходные или выделенные информативные параметры имеют неравнозначную важность для определения структуры объектов. Для выявления таких показателей важности в работе предлагается использовать процедуры экспертного оценивания. Наиболее хорошие результаты дает процедура многовариантной экспертизы [5], когда для получения параметра важности для каждого оцениваемого параметра используется несколько групп экспертов – специалистов в

различных аспектах исследуемой проблемы. В результате экспертизы каждый параметр получает определённый вес (показатель «важности» этого параметра) для формирования структуры объектов.

1.3. **Алгоритм построения начального разбиения.** На первом шаге из всех точек выборки $x_1, \ldots, x_n$ находится пара наиболее удаленных друг от друга точек, x_l и x_p, одна из которых – x_l, относится к первому классу, а другая x_p – ко второму. Если n достаточно велико, то используется усеченный вариант первого шага, а именно: x_l выбирается случайно, а x_p ищется как точка, наиболее от неё удаленная.

На втором шаге ищутся точки x_{l+1} и x_{p+1} – ближайшие, соответственно, к точкам x_l и x_p; точка x_{l+1} относится к первому классу, а x_{p+1} – ко второму.

На (s+1)-ом шаге ищутся точки x_{l+s} и x_{p+s}, ближайшие в среднем к уже найденным точкам, соответственно, первого и второго классов. Точка x_{l+s} определяется следующим образом:

$$x_{l+s} = x_j / \min_{x_j} \frac{1}{s} \sum_{m=0}^{s} K(x_j, x_{l+m}). \tag{5}$$

Точка x_{p+s} определяется аналогично. Если возникает «конфликт», т.е. одна и та же точка является ближайшей к первому и ко второму классам одновременно, то эта точка относится к первому классу. Процедура (5) повторяется до тех пор, пока не будут исчерпаны все точки выборки. Полученное разбиение принимается в качестве начального разбиения R_0.

1.4. **Алгоритм выбора числа классов.** Для выбора числа классов используется специальная экспертно-компьютерная процедура, которая работает следующим образом. Сначала эксперт-пользователь оценивает диапазон $(r_{\min}, r_{\max})$, в пределах которого заведомо находится искомое число классов. Далее, используя любой алгоритм автоматической классификации (в настоящей работе применялся алгоритм m-локальной оптимизации), проводится разбиение анализируемого множества объектов на $r_{\min},\ r_{\min}+1,\ \ldots,\ r_{\max}$ классов. Качество каждой из полученных классификаций оценивалось с помощью критерия $J_3 = J_1 - qJ_2$, где J_1 вычисляется по формуле (3), J_2, а также некоторые вспомогательные величины вычисляются по формулам:

$J_2 = \frac{1}{r-1} \sum_{i=1}^{r} \sum_{j>i} \frac{n_i+n_j}{n} K(A_i, A_j)$; $K(A_i, A_j) = \frac{1}{n_i\, n_j} \sum_{x_l \in A_i} \sum_{x_p \in A_j} K(x_l, x_p)$ – мера близости классов $A_i,\ A_j$; где потенциальная функция $K(x_i, x_j)$ определяется формулой (1); q, α и p из (1) – настраиваемые параметры алгоритма. Фактически, параметр q является масштабирующим параметром, приводящим к соизмеримым средним значениям функционалов J_1 и J_2; на практике величина q имеет значение порядка 2-7 (обычно во столько раз отличается средняя близость внутри классов от средней близости между самими классами).

Формально, в качестве «оптимального» можно выбрать такое число классов r_{opt}, которое соответствует максимальному значению $J_3(r_j)$, т.е. $r_{opt} = r_j$, для которого $\max J_3(r_j), \quad r_j = r_{\min}, \ldots, r_{\max}$. Однако наличие существенной, но неиспользованной при классификации информации, например, ввиду отсутствия данных, может привести к тому, что полученное таким способом r_{opt} не будет «истинно оптимальным».

Для компенсации этого недостатка предлагается использовать следующую экспертную процедуру. Экспертам – специалистам в соответствующей предметной области представляются значения $J_3(r_j)$, $r_j = r_{\min}, \ldots, r_{\max}$, представленные для удобства в виде графика, на котором отмечается значение r_{opt} (оно соответствует максимальной точке на графике $J_3(r_j)$). Используя эту информацию, эксперты могут корректировать выбираемое число классов. В подавляющем числе случаев экспертное число классов либо совпадает с r_{opt}, либо незначительно (± 1) отличается от него.

При классификации многомерных объектов во время такой экспертизы анализируется также классификация каждого объекта. Для этой цели экспертам сообщается информация о мере близости $K(x_i, c_j)$ каждой точки x_i до центров классов $c_j \quad j = 1, \ldots, r_{opt}$ в оптимальной классификации, т.е. матрица близости $\|K(x_i, c_j)\|$, $i = 1, \ldots, n$, $j = 1, \ldots, r_{opt}$. Перенесение точки (объекта) x_i из j-го класса в l-й считается допустимым, если величины $K(x_i, c_j)$ и $K(x_i, c_l)$ отличаются незначительно. Другими словами, содержательно обоснованное перенесение допустимо для точек, расположенных вблизи границы между соответствующими классами.

1.5. **Алгоритм заполнения пропущенных наблюдений.** Во многих приложениях имеются пропуски в данных. В этой ситуации нужно либо использовать специальные процедуры подсчета расстояний между объектами, в параметрах которых имеются пропуски, либо разрабатывать специальные процедуры заполнения таких пропусков. В подавляющем большинстве случаев, пропуски по каждому параметру заполняются средним известных значений соответствующего параметра (для исходной выборки). В настоящей работе была разработана специальная процедура заполнения пропусков в исходных данных с использованием алгоритмов автоматической классификации. Основная идея процедуры состоит в следующем. Если множество изучаемых объектов структурировано (т.е. их можно разделить на классы, достаточно компактно расположенные в пространстве параметров X), то дисперсия (диапазон) изменения каждого параметра в пределах каждой группы, как правило, будет существенно меньше, чем этот показатель для значения этого параметра по всей выборке. Таким образом, если по данным с пропусками удастся определить реальную структуру взаиморасположения точек (т.е. провести классификацию, адекватную этой структуре), то заполнять пропущенное значение l-го параметра для объекта из i-го класса можно средним этого параметра по его известным значениям для всех объектов, попавших в i-ый класс. Исходя из сделанного предположения, отклонение полученного значения от «истинного» должно быть существенно меньше (в среднем), чем обычная схема заполнения по общему среднему.

Заключение

Разработанный комплексный алгоритм использовался для интеллектуализации анализа сложноорганизованных данных, а также при совершенствовании процедур принятия решений для нескольких крупных систем управления, в основном регионального характера. Во всех приложениях, а также при машинном моделировании, была подтверждена высокая эффективность разработанного комплексного алгоритма.

Список литературы

1. *Бауман Е.В., Дорофеюк А.А.* Классификационный анализ данных / Труды Международной конференции по проблемам управления. Том 1. – М.: СИНТЕГ, 1999. – С. 62-67.
2. *Дорофеюк А.А., Дорофеюк Ю.А* Методы структурно-классификационного прогнозирования многомерных динамических объектов / Искусственный интеллект, № 2, 2006. – С.138-141.
3. *Браверман Э.М., Мучник И.Б.* Структурные методы обработки эмпирических данных – М.: Наука, 1983.
4. *Десова А.А., Дорофеюк А.А., Гучук В.В., Дорофеюк Ю.А., Покровская И.В.* Процедуры классификационного анализа в задаче формирования информативных признаков при исследовании ритмической структуры биосигнала / Автоматика и телемеханика. 2008, №6. – С. 143-152.
5. *Дорофеюк А.А., Покровская И.В., Чернявский А.Л.* Экспертные методы анализа и совершенствования систем управления / Автоматика и телемеханика. 2004, №10. – С. 172-188.

Статья поступила в редакцию 27.04.2008