УДК 519.72 *МРНТИ 27.47.17*

А. А. Зейнуллина, Ж. А. Масимханова, С. А. Мустафин, к.т.н.

Институт проблем информатики и управления

АНАЛИЗ МНОГОМЕРНЫХ ДАННЫХ В ЗАДАЧАХ ОПТИМИЗАЦИИ

Рассмотрены проблемы оптимизационных задач, связанные с анализом данных. Предложен метод оптимизации функции от нескольких переменных, основанный на методе автоматической классификации.
Ключевые слова: анализ, многомерные данные, оптимизация функции, автоматическая классификация.

Мақалада деректерді талдаумен байланысты оңтайландыру міндеттерінің проблемалары карастырылды. Автоматты түрде классификациялау әдісіне негізделген бірнеше айнымалылардан функцияны оңтайландыру әдісі ұсынылды.
Кілт сөздер: талдау, көп өлшемді мәліметтер, функцияны оңтайландыру, автоматты классификация.

The article discusses the problems of optimization tasks that associated with the analysis of data. A method for optimizing a function of several variables, based on the method of automatic classification is proposed.
Key words: analysis, multidimensional data, optimization of functions, automatic classification.

Практические приложения способствовали прогрессу в развитии алгоритмов обработки многомерных данных и, как следствие, увеличению объема информации, времени ее обработки и сложности производимых операций. Применение традиционных вычислительных средств для обработки данных в реальном масштабе времени оказалось затруднительно и требует использования специальных средств обработки. Оптимизационный анализ основан на массовом решении обратных задач

при изменяющихся в определенных интервалах параметрах рассматриваемого класса задач. Для этого необходима общая схема работы с подобными данными.

Пусть даны некоторые объекты $x_1,\ldots,x_m$, принадлежащие пространству объектов X. Истинная классификация этих объектов неизвестна, их необходимо классифицировать, разбив исходное множество на классы таким образом, чтобы в каждом классе оказались включенными объекты, близкие между собой в том или ином смысле.

Построение классификации объектов $x_1,\ldots,x_m$ формализует интуитивное понятие "схожести" объектов. И это понятие может пониматься по-разному. Поэтому возможны различные пути его формализации. Обычно в автоматической классификации рассматривают такое понимание "сходства", которое вытекает из геометрических представлений об объектах как точках в пространстве объектов X. Отсюда будет вытекать соответствующая формализация этого понятия. Конкретно мы будем считать, что для любых двух объектов x_1, x_2 определено расстояние между ними $\rho(x_1, x_2)$. Отметим, что выполнение аксиомы треугольника необязательно. Естественно считать два объекта тем более "схожими", чем ближе они находятся друг к другу в смысле расстояния ρ. Вводя ρ, мы тем самым неявно готовимся к формированию искомых классов. Очевидно, что такая интерпретация "схожести" будет эффективна только в том случае, когда введенное расстояние ρ соответствует содержательной стороне конкретной рассматриваемой задачи классификации. Выбор расстояния для каждой конкретной задачи классификации является неформальной процедурой и осуществляется проектирующим систему.

Определение "схожести" между объектами с помощью расстояния между ними в пространстве объектов X приводит к тому, что "схожими" будут те объекты, которые собраны в компактные группы. Каждая такая группа объектов и образует класс, т. е.

рассматриваемая интерпретация "схожести" имеет смысл при истинности гипотезы компактности объектов.

Такой подход к определению "схожести" ориентирован на использование информации о классах, содержащейся во взаимном расположении объектов в пространстве объектов X. В связи с этим алгоритм, реализующий решение задачи классификации, должен уметь выделять в пространстве X области с большой плотностью объектов из последовательности $x_1,\ldots, x_m$ [1].

Пусть дана функция $F = F(Z)$ от n переменных $z_1,\ldots, z_n$. Переменные принадлежат области

$$P = (z_i \mid a_i \le z_i \le b_i, \quad i = 1,n) \tag{1}$$

Переменные $z_i, \quad i = 1,\ldots, n$ могут быть связаны некоторыми соотношениями:

$$D = (Z \mid \quad R_k(Z) \ge 0, \quad k = 1,\ldots,l) \tag{2}$$

В области $P \cap D$ требуется найти точку $Z^* = (z_1^*,\ldots z_n^*)$, доставляющую глобальный минимум функции $F = F(Z)$:

$$Z^* = \arg\min\{ F(Z) \mid Z \in P \bigcap D\}.$$

Большинство методов нахождения глобального экстремума функции нескольких переменных основано на поиске локальных экстремумов. Неотъемлемой частью задачи автоматической классификации является введение понятия оптимального критерия, которое позволяет установить, когда достигается желательное разделение. Для введения подобного критерия необходимо найти меру внутренней однородности класса и меру разнородности классов между собой. Это понятие достаточно размытое и исследователями понимается по-разному. Для части алгоритмов автоматической классификации такой функционал подразумевается как существующий. Для других алгоритмов автоматической классификации такой функционал следует из самой постановки конкретной задачи или является требованием заказчика. Кроме того, алгоритмы решения задачи автома-

тической классификации содержат неявно в своем построении функционал качества.

Построение классификации приводит к ряду "плотных" классов, которые соединяются "неплотными" классами - мостиками. Желательно при этом пользоваться методами, которые бы определили моды этого распределения и соответствующие им отдельные классы. Для этого рассмотрим следующий метод классификации. Этот метод начинается с выяснения вопроса о мультимодальности данных. В случае одного признака необходимо построить гистограмму и вычеркнуть данные с малой частотой (седловые области). Тогда соответствующий класс можно установить для каждой модальной области. Данные, принадлежащие седловой области, относят к ближайшей моде. В случае n, $n \succ 1$ признаков этот метод становится неудобным. Тогда алгоритм классификации может быть записан следующим образом:

1. Выбираем значение радиуса r гиперсферы (окрестности).

2. Центр $C^{(1)}$ окрестности $O(C^{(1)},r)$ совмещаем с любой точкой исходного множества точек.

3. Определяем точки, попавшие в окрестность $O(C^{(1)},r)$.

4. Центр окрестности $O(C^{(1)},r)$ смещаем в точку $C^{(2)}$ - среднюю точек, попавших на предыдущем шаге в окрестность $O(C^{(1)},r)$

5. Процедура продолжается до тех пор, пока не перестанут изменяться координаты средних точек $C^{(i)}$.

Очевидно, при этом окрестность остановится в области локального максимума плотности точек. После остановки точки, попавшие в последовательность окрестностей, исключаются. Затем центр окрестности совмещается с любой из точек оставшегося множества точек, и процедура повторяется до тех пор, пока все исходное множество точек не будет разделено на классы $K_i, i = 1, q$ [1]. В итоге получаем набор классов $K_i, i = 1, q$, каж-

дый из которых может быть представлен множеством центров окрестностей радиуса r.

Поисковые методы решения многомерных оптимизационных задач сводят задачу к системе локальных подзадач и применяют методы локальной оптимизации типа градиентного спуска. Локальные методы осуществляют такое сведение путем построения направлений спуска, вдоль которых осуществляется локальная минимизация, что порождает траекторию, ведущую из начальной точки в окрестность решения.

В предлагаемом алгоритме вводится новая система ограничений, основанная на разбиении области допустимых значений переменных $P \cap D$ и исключающая возможность вторичного попадания в найденные классы. Введение этих ограничений позволяет по-новому решать и проблему окончания поиска глобального минимума.

Идея поиска минимум $F(Z)$ заключается в рассмотрении соответствия классификации значений целевой функции $F(Z)$ в задаваемых точках $P \cap D$ и классификации этих точек [2]. Таким образом, до применения трудоемких алгоритмов оптимизации выясняется, как распределяется данная совокупность точек из $P \cap D$ на классы в зависимости от значений $F(Z)$ - по заданной функции и последовательности точек из $P \cap D$ находят разбиение $P \cap D$.

Пусть область $P \cap D$ разбита на L классов $K_1,..., K_L$ в соответствии с разбиением значений целевой функции $F = F(Z)$ на классы в точках $z^1,..., z^M \in P \cap D$. Далее в каждой полученной области $P \cap D$, представленной точками класса $K_i, i = 1,..., L$, производится поиск локального минимума:

$$z_i^* = \arg\min\{ F(z) \mid z \in K_i \}, i = 1,..., L.\}.$$

Можно считать, что разделяющая классы поверхность строится сначала для отделения точек $P \cap D$, которые участвовали в поиске первого локального минимума. Затем процедура по-

иска повторяется на множестве оставшихся точек из области $P \cap D$ для выделения второго класса и т. д.

Другими словами, после нахождения очередного класса K_j и соответствующего ему локального минимума S_j часть $P \cap D$, которая представлена точками K_j и точками $P \cap D$, приводящими к соответствующей точке S_j (класс $\tilde{K}_j$), объявляются запретными областями $\tilde{K}_j$ для входа новых точек поиска. Если при некотором спуске конечная точка T очередного локального поиска "наткнулась" на область $\tilde{K}_j$, то радиус окрестности точки $H \in \tilde{K}_j$, ближайшей к точке T, увеличивается на величину ΔR_H ($\Delta R_H > 0$):

$$\tilde{K}_j = \tilde{K}_j \cup O(R_H + \Delta R_H),$$

где $O(R_H)$ - окрестность точки $H \in \tilde{K}_j$ радиуса R_H.

Тогда глобальный поиск проводится или до выполнения ограничений на число попаданий в запретные классы, или до полного покрытия области $P \cap D$ расширенными окрестностями, или по точности, или отказом от поиска. Данный подход проходил апробацию на классических тестах, некоторых практических задачах и показал неплохие результаты.

Литература

1 *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. Основы моделирования и первичная обработка данных. - М., 1983. - Т.1. - 472 с.

2 *Мустафин С. А.* Методы построения скелетов изображений // Новости науки Казахстана. - 2002. - Вып.1. - С. 24-26.