

## ОПТИМИЗАЦИЯ И ЗАДАЧА АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ

Рассматривается метод оптимизации функции от нескольких переменных, отличающийся от предлагаемых методов введением дополнительной системы классов-ограничений, что исключает возможность попадания в найденные минимумы, и условием окончания глобального поиска.

Пусть даны некоторые объекты  $x_1, \dots, x_m$ , принадлежащие пространству объектов  $X$ . Истинная классификация этих объектов неизвестна. Необходимо классифицировать эти объекты – разбить их исходное множество на классы таким образом, чтобы в каждом классе оказались включенными объекты, близкие между собой в том или ином смысле. Априорная информация о классификации отсутствует. Это постановка задачи автоматической классификации.

Построение классификации объектов  $x_1, \dots, x_m$  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  $\rho$ . Вводя  $\rho$ , мы тем самым неявно готовимся к формированию искомых классов. Очевидно, что такая интерпретация «схожести» будет эффективна только в том случае, когда введенное расстояние  $\rho$  соответствует содержательной стороне конкретной рассматриваемой задачи классификации. Выбор расстояния для каждой конкретной задачи классификации является неформальной процедурой и осуществляется обычно лицом, проектирующим систему классификации.

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 один класс (т.е. все эти объекты принадлежат одному образу). Это означает также, что рассматриваемая интерпретация «схожести» имеет смысл при предположении о том, что гипотеза компактности объектов имеет место.

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  $x_1, \dots, x_m$  и игнорировать области с малой плотностью [1].

Рассмотрим постановку задачи оптимизации. Пусть дана функция  $F = F(Z)$  от  $n$  переменных  $z_1, \dots, z_n$ . Переменные принадлежат области

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

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

$$D = \{Z \mid R_k(Z) \geq 0, \quad k = 1, \dots, l\}. \quad (2)$$

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

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

Большинство методов нахождения глобального экстремума функции нескольких переменных основано на поиске локальных экстремумов. Проблемой при этом является заикливание алгоритмов, невозможность выйти из «оврага» и другие причины. Для выхода из такой ситуации применяют ряд приемов.

Связь задач оптимизации и классификации. Неотъемлемой частью задачи автоматической классификации является введение понятия оптимального критерия (целевой функции), 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, по аналогии существования разделяющих поверхностей в распознавании образов. Для другой части алгоритмов автоматической классификации такой функционал следует из самой постановки конкретной задачи или является требованием заказчика. Кроме того, алгоритмы решения задачи автоматической классификации содержат неявно в своем построении интерпретирующий качество классификации функционал. Заметим, что приведенное деление задач автоматической классификации по функционалам качества разбиения данных является условным.

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  $n, n \neq 1$ , 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 двигающихся окрестностей, из дальнейшего рассмотрения 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  $K_i, i = 1, q$ , каждый из которых может быть представлен множеством центров окрестностей радиуса  $r$ .

Предлагаемый алгоритм оптимизации. Поиск методы решения многомерных оптимизационных задач, как уже отмечено, 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 (области локальных минимумов). Введение этих ограничений позволяет по-новому решать и проблему окончания поиска глобального минимума.

Идея поиска минимума  $F(Z)$  заключается в рассмотрении соответствия классификации значений целевой функции  $F(Z)$  в задаваемых точках  $P \cap D$  и классификации этих точек [2]. Другими словами, до применения тех или иных, нередко 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

$P \cap D$ , т.е. происходит как бы подстройка разбиения  $P \cap D$  под функцию  $F(Z)$ .

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

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

Можно считать, что разделяющая классы поверхность, строится сначала для отделения точек  $P \cap D$ , которые участвовали в поиске первого локального минимума. Далее, процедура поиска повторяется на множестве оставшихся точек из области  $P \cap D$  для выделения второго класса, точки которого приводят ко второму локальному минимуму 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  $\Delta R_H$  ( $\Delta R_H \neq 0$ ):

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

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 оптимизации.

Этот метод был реализован программно на ПК, проходил апробацию на известных примерах, приведенных в литературе по оптимизации, и показал неплохие результаты. Данный метод был применен для решения одной задачи по определению оптимальных параметров технологического процесса добычи руд и его результаты удовлетворили заказчиков.

#### ЛИТЕРАТУРА

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. т.1. Основы моделирования и первичная обработка данных. М., 1983. 472 с.
2. Айдарханов М.Б. О некоторых метрических свойствах пространства классификаций // Журнал вычислительной математики и математической физики. 1991. Т. 31. С. 169-173.
3. Мустафин С.А. Методы построения скелетов изображений // Новости науки Казахстана. Алматы, 2002. Вып. 1. С. 24-26.

#### Резюме

Ұсынылған мақалада бірнеше айнымалыға тәуелді функцияның тиімді әдісі ұсынылады. Бұл әдістің алдыңғы ұсынылған әдістерден ерекшелігі өз бетінше оқу түрі, қосымша шектеулер кластар жүйесін енгізу және глобалды іздеуді аяқтаудың шартында болып табылады.

#### Summary

In this paper the method of optimization of function from several variables is proposed. Its difference from a earlier offered methods consists in a type of learning, in introduction of additional system of restrictions and in condition 's end of global search.

Институт информатики

и управления МОН РК, г. Алматы Поступила 30.04.07г.