ՀԱՅԿԱԿԱՆ ՍՍՀ ԳԻՏՈՒԹՅՈՒՆՆԵՐԻ ԱԿԱԴԵՄԻԱՅԻ ՏԵՂԵԿԱԳԻՐ
ИЗВЕСТИЯ АКАДЕМИИ НАУК АРМЯНСКОЙ ССР

Տեխնիկական գիտութ. սերիա XXXV, № 2, 1982 Серия технических наук

ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

Ю. М. ГАСПАРЯН, В. М. МОВСЕСЯН, А. Л. НЕРСЕСЯН

К ВОПРОСУ ВЫБОРА ОПТИМАЛЬНОГО ЗНАЧЕНИЯ ПОРОГА В ЗАДАЧАХ КЛАССИФИКАЦИИ

В некоторых алгоритмах классификации качество контролируется с помощью некоторого функционала [1—3] и ищется оптимальная классификация, которая бы максимизировала (или минимизировала) этот функционал. При большом количестве классифицируемых объектов число рассматриваемых вариантов разбиения на классы резко возрастает, что приводит к большим затратам времени на классификацию. Поэтому целесообразно уменьшить число рассматриваемых вариантов путем исключения из рассмотрения тех вариантов разбиения, которые заранее выглядят нецелесообразными. Для этого в некоторых алгоритмах классификации меры близости между объектами сравниваются с порогом значимости T. Если меру близости между парами объектов X_i и X_j ($i=\overline{1, N}$, $j=\overline{1, N}$) представить как элементы матрицы связей $A=(a_{ij})_{N\times N}$, то после сравнения с порогом T из матрицы A формируется новая матрица $D=(d_{ij})_{N\times N}$, состоящая из нулей и единиц [4], где

$$d_{ij}=\begin{cases}1, & \text{если } a_{ij}>T,\\ 0, & \text{если } a_{ij}\leqslant T.\end{cases} \qquad (1)$$

Для классификации можно воспользоваться матрицей D. Благодаря тому, что часть элементов матрицы D равна нулю, число возможных вариантов разбиений существенно сокращается по сравнению с разбиением с использованием матрицы A.

Очевидно, что степень заполненности матрицы D зависит от значения порога значимости T. В случае, когда меры близости нормированы в интервале $[0, 1]$, при значении порога значимости $T=0$ все объекты попарно считаются похожими, т. е. матрица состоит из одних единиц, и каждый объект может войти в любой класс. В случае, когда $T=1$, все недиагональные элементы матрицы D равны нулю и объекты попарно не похожи — каждый объект составляет самостоятельный класс. При промежуточных значениях $T\in[0, 1]$ матрица D заполняется единицами частично. Следовательно, при больших значениях T матрица D будет слабо заполненной, т. е. будут преобладать нулевые эле-

менты, что уменьшит число рассматриваемых вариантов и приведет к упрощению процедуры классификации. При небольших значениях T в матрице D преобладают единицы и процедура классификации усложняется. Следует иметь в виду также, что при сравнении мер близости a_{ij} с порогом, т. е. при аппроксимации матрицы A матрицей D теряется некоторая информация, содержащаяся в матрице A. При этом, чем ближе значение T к единице, тем ошибка аппроксимации будет больше.

Множество объектов и взаимосвязь между объектами можно представить в виде взвешенного графа G, вершинами которого являются объекты $x_1, \ldots, x_N$, а весами ребер (x_i, x_j), $(i = \overline{1, N};\ j = \overline{1, N})$ — элементы a_{ij}, $(i = \overline{1, N};\ j = \overline{1, N})$ матрицы A. После сравнения элементов матрицы A с порогом, она аппроксимируется D, которая представляет собой матрицу инцидентности невзвешенного помеченного графа Г. Если $d_{ij} = 1$, то существует ребро между вершинами X_i и X_j, и если $d_{ij} = 0$, то отсутствует ребро, соединяющее вершины X_i и X_j. Сравнение мер близости a_{ij} с порогом T фактически аппроксимирует взвешанный граф G невзвешанным, помеченным графом Г.

Классификация на основе графа Г (матрицы D) тем труднее, чем больше: а) количество вершин графа (количество объектов); б) число ребер в графе (число единиц в матрице D); в) общее количество путей, соединяющих пары вершин.

Перечисленные условия, которые характеризуют трудности проведения классификации, можно объединить в одну характеристику графа Г, с помощью понятия комбинаторной сложности графа, введенной в [5]. Функция комбинаторной сложности графа представляется формулой [5]:

$$\chi(\Gamma) = \frac{N \cdot M}{N + M} \sum_{\substack{(i,j) \\ i<j}} \sum_{k=1}^{n} \gamma_k(i, j), \qquad (2)$$

где N, M — число вершин и ребер графа; $\gamma_k(i, j)$ — число путей длины K из вершины X_i к вершине X_j. Функцию сложности (2) можно представить и в другом виде:

$$\chi(\Gamma) = \frac{N \cdot M}{N + M} \sum_{\substack{(i,j) \\ i<j}} s_{ij}, \qquad (3)$$

где s_{ij} — общая суммарная длина путей, соединяющих пару вершин i и j.

Очевидно, что значение функции сложности χ (Г) графа Г зависит от значения порога значимости T. Для упрощения процедуры классификации необходимо выбрать T таким образом, чтобы функция (2) принимала минимальное значение, а от значения порога T зависит и ошибка аппроксимации.

В некоторых алгоритмах классификации в качестве меры близости между двумя объектами используется передача информации $I(X_i, X_j)$, но при этом не учитываются информационные взаимосвязи между тремя и большим числом объектов. В этом случае элементами матрицы связей A являются передачи информации между двумя объектами $a_{ij} = I(X_i, X_j)$; $a_{ij} = a_{ji}$. После сравнения с порогом, передачам информации между двумя объектами присваиваются значения либо «1», либо «0» в соответствии с (1). Потерю информации вследствии аппроксимации можно количественно оценить величиной

$$\Delta I = \left| \sum_{\substack{(i,j) \\ i>j}} I(X_i, X_j) - \sum_{\substack{(i,j) \\ i>j}} d_{ij} \right|. \qquad (5)$$

Теперь задача определения оптимального значения порога значимости T сводится к определению такого значения T, при котором комбинаторная сложность графа Γ принимала бы минимальное значение, а потеря информации (5) не превышала заданное значение. Математически это можно представить в виде задачи математического программирования, т. е.

$$\min_{T} \chi(\Gamma),$$

$$\Delta I \leqslant \Delta I_{\text{доп}}. \qquad (6)$$

Задачу (6) можно решать поисковыми алгоритмами.

С ростом порога значимости T убывает число единиц в матрице D, следовательно, уменьшаются числа ребер и путей в графе Γ. Поэтому функция сложности $\chi(\Gamma)$ является монотонно невозрастающей относительно порога значимости T, т. е. при $\Delta T > 0$, $\Delta\chi(\Gamma) \leqslant 0$. Функция потери информации (5) состоит из разности двух членов, один из них $\sum I(X_i X_j)$ имеет постоянное значение, а второй — $\sum d_{ij}$ — с ростом порога T монотонно убывает. Следовательно, пока $\sum I(X_i X_j) < \sum d_{ij}$, потеря информации ΔI с ростом T убывает. При некотором значении порога T_m, ΔI имеет минимальное значение. Начиная с этого значения рост порога T сопровождается монотонным ростом ΔI. Таким образом, при изменении T от 0 до T_m функция ΔI монотонно не возрастает, а при изменении T от T_m до 1 ΔI монотонно не убывает, т. е. является унимодальной функцией от T.

Из свойств функций $\chi(\Gamma)$ и ΔI следует, что для решения задачи (6) можно применить итерационные алгоритмы градиентного типа.

Построим функцию Лагранжа:

$$\Phi(T, \lambda) = \chi(\Gamma) + \lambda\Theta(T), \qquad (7)$$

где λ — неопределенный множитель Лагранжа, а $\Theta(T) = \Delta T_{\text{доп}} - \Delta I$.

Решение задачи (6), которое соответствует минимуму функции (7), определяется итерационной процедурой вида:

$$T^{n+1} = T^n - \alpha \left| \frac{\chi^n(\Gamma) - \chi^{n-1}(\Gamma)}{T^n - T^{n-1}} + \varepsilon^n \frac{\Theta^n - \Theta^{n-1}}{T^n - T^{n-1}} \right|;$$

$$\varepsilon^{n+1} = \max\{0,\ \varepsilon^n + \alpha\Theta^n\}, \qquad (8)$$

$$\alpha > 0,$$

где α — параметр, характеризующий размер шага, а $\chi^n(\Gamma)$, Θ^n, ε^n — значения соответствующих величин при значении порога T^n. Значения $\chi(\Gamma)$ и Θ, как функции порога T, имеют ступенчатый характер, поэтому итерационная процедура (8) при малых изменениях T может остановиться в пологих областях функций $\chi(\Gamma)$ и Θ. Во избежание этого, функции $\chi(\Gamma)$ и Θ можно сглаживать операторами, либо сделать шаг α в процедуре (8) переменным.

При большом количестве классифицируемых объектов, вычисление комбинаторной сложности формулами (2) и (3) трудоемко, поэтому при больших значениях N можно пользоваться оценками функции сложности, приведенными в [5].

ЕрПИ им. К. Маркса

Поступило 26.V 1981

Эм. Г. ԳԱՍՊԱՐՅԱՆ, Վ. Ա. ՄՈՎՍԻՍՅԱՆ, Ա. Լ. ՆԵՐՍԻՍՅԱՆ

ՇԵՄՔԻ ՕՊՏԻՄԱԼ ՄԵԾՈՒԹՅԱՆ ԸՆՏՐՈՒԹՅԱՆ ՀԱՐՑԻ ՎԵՐԱԲԵՐՅԱԼ ԳՈՒՄԱԿԱՐԳՄԱՆ ԱՆԴԻՐՆԵՐՈՒՄ

Ա մ փ ո փ ո ւ մ

Դիտարկվում է պատկերման խնդիրներում շեմքի օպտիմալ մեծության ալգորիթմ։ Այդ խնդրի լուծման համար հաշվի է առնվում պատճի կամքինաստոր բարդություններ։ Հոդվածում բերվում է շեմքի օպտիմալ մեծության ընտրման ալգորիթմը, որը ապահովում է նվազագույն քարդություն տրված սխալի մակարդակի դեպքում։ Դիտարկված ալգորիթմը կարելի է օգտագործել պատկերների Տանուշման, նույնականացման ժամանակ խնֆորմատիկ պարամետրերի ընտրման խնդիրներում և այլն։

ЛИТЕРАТУРА

1. *Браверман Э. М.* и др. Диагонализация матрицы связи и выявление скрытых факторов. Сб. «Проблемы расширения возможностей автоматов», вып. 1, М., «Наука», 1971.

2. *Куперштох В. Л., Миркин Б. Г., Трофимов В. А.* К обоснованию одного критерия классификации. Сб. «Методы моделирования и обработка информации», Новосибирск, Сиб. отд. изд. «Наука», 1976.

3. *Koontz Warren L. C., Narendra Patrenahalli M., Fukunaga Reinosuke.* A branch and Bound clustering algorithm. IEEE Trans. Comp., v. 24, № 9, 1975.

4. *Бонер Р. Е.* Некоторые методы классификации. Сб. пер. «Автоматический анализ сложных изображений», под. ред. Э. М. Бравермана, М., «Мир», 1969.

5. *Minoli D.* Combinatorical graph complexity. "Atti della Accademia nazionale dei Lincei, Rendiconti classe di Scienze fisiche, matematiche e naturali", v. 59, № 6, 1975.