УДК 004.852

Об оптимизации билинейных форм в задачах классификации

В.В. Рязанов[1]

[1]Московский физико-технический институт (государственный университет)

В работе предлагается один алгоритм оптимизации билинейных форм в теории классификации. Подобные задачи возникают в ряде моделей, когда требуется выделение «информативной» подтаблицы таблицы обучения, то есть одновременное выделение в таблице обучения информативных признаков и объектов. Так в моделях вычисления оценок в одном из способов вычисления оценок объектов за классы оценки вычисляются как линейные функции параметров $\mathbf{x} = (x_1, x_2, ..., x_n)$ «веса признаков» и одновременно линейные от параметров, характеризующих «веса объектов» - $\mathbf{y} = (y_1, y_2, ..., y_m)$. В качестве функционала, характеризующего точность распознавания контрольной выборки, можно использовать эвристический функционал $f(\mathbf{x}, \mathbf{y}, t)$ - разность неотрицательных нормированных сумм оценок объектов за свои классы и аналогичной суммы оценок объектов за чужие классы, помноженной на скалярный неотрицательный параметр t. Данный параметр нужен для установления оптимального соответствия в функционале $f(\mathbf{x}, \mathbf{y}, t)$ между оценками за «свой» класс и оценками за «чужие» классы.

В алгоритмах АВО [1, 2] оценка $\Gamma_j(\mathbf{z})$ распознаваемого объекта $\mathbf{z} = (z_1, z_2, ..., z_n)$ к классу K_j выражается как $\Gamma_j(\mathbf{z}) = \frac{1}{|K_j|} \sum_{\mathbf{z}_\nu \in K_j} y_\nu \sum_{\Omega \in \Omega_A} (\sum_{i \in \Omega} x_i) B_\Omega(\mathbf{z}_\nu, \mathbf{z})$, где $|K_j|$ - число эталонных объектов $\mathbf{z}_\nu$ в классе K_j, Ω - некоторое опорное множество (подмножество признаков) из совокупности опорных множеств Ω_A алгоритма АВО, $B_\Omega(\mathbf{z}_\nu, \mathbf{z})$ - близость объекта $\mathbf{z}$ к объекту обучения $\mathbf{z}_\nu$ по опорному множеству. Можно показать, что оценка $\Gamma_j(\mathbf{z}) = \frac{1}{|K_j|} \sum_{\mathbf{z}_i \in K_j} y_i (\sum_{t \in J(\mathbf{z}, \mathbf{z}_i)} x_t) C^{k-1}_{d(\mathbf{z}, \mathbf{z}_i)-1}$, где $J(\mathbf{z}, \mathbf{z}_i) = \{\nu : |z_{i\nu} - z_\nu| \le \varepsilon_\nu, \nu = 1, ..., n\}$, $d(\mathbf{z}, \mathbf{z}_i) = |J(\mathbf{z}, \mathbf{z}_i)|$, $k, \varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ - некоторые параметры алгоритма распознавания. Здесь в качестве опорных множеств используются всевозможные подмножества из k признаков. Тогда $f(\mathbf{x}, \mathbf{y}, t) = \sum_{\alpha=1}^{l} \sum_{\mathbf{z}'_\tau \in K_\alpha} \Gamma_\alpha(\mathbf{z}'_\tau) - t \sum_{\alpha=1}^{l} \sum_{\mathbf{z}'_\tau \notin K_\alpha} \Gamma_\alpha(\mathbf{z}'_\tau)$ (где $\mathbf{z}'_\tau$ - элементы контрольной выборки), что является билинейной формой от параметров $x_1, x_2, ..., x_n, y_1, y_2, ..., y_m$.

Итак, рассматривается следующая оптимизационная задача

$$f(\mathbf{x}, \mathbf{y}, t) = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij}(t) x_i y_j \to \max_{\mathbf{x} \in X, \mathbf{y} \in Y} \qquad (1)$$

Для выбора областей X и Y используются два варианта: $1 \ge x_i \ge 0, i = 1, 2, ..., n, 1 \ge y_j \ge 0, j = 1, 2, ..., m,$ или $x_i \in \{0, 1\}, i = 1, 2, ..., n, y_j \in \{0, 1\}, j = 1, 2, ..., m$. В последнем случае выбор единичных значений параметров $x_1, x_2, ..., x_n, y_1, y_2, ..., y_m$ означает выбор некоторой подтаблицы таблицы обучения.

Для решения задачи (1) для случая дискретной оптимизации разработан (и исследован на модельных данных и данных из репозитория [3]) приближенный локальный метод, имеющий линейную сложность при малых окрестностях поиска.

Исследована также модификация модели поиска оптимальных $t, \mathbf{x}, \mathbf{y}$. Представляет интерес задача поиска оптимального значения параметра t, при котором будет наилучшее распознавание. Был использован стандартный критерий качества распознавания (доля правильно распознанных объектов контрольной выборки при выбранной подтаблице обучения) $F(\mathbf{x}, \mathbf{y}, t)$ и создан метод его максимизации.

Работа выполнена при поддержке гранта РФФИ № 16-31-00443.

Литература

1. *Журавлев Ю.И., Никифоров В.В.* Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. 1971. №3. С. 1-11.
2. *Журавлев Ю.И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. Вып. 33. М.: Наука, 1978. 5-68 с.
3. *Lichman, M*. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml] // Irvine, CA: University of California, School of Information and Computer Science. 2013.