Способы построения оптимальной вероятностной модели систем распознавания

Капустий Б. Е., Русын Б. П., Таянов В. А.
vtayanov@ipm.lviv.ua
Украина, Львов, Физико-механический институт им. Г. В. Карпенка НАН Украины

Актуальность задачи построения математической модели систем распознавания (СР) состоит в том, что она позволяет исследовать эту систему, не реализуя её в полном объёме. Определение параметров проводится на основании обучающей выборки. Оптимальность модели определяется её точностью и скоростью вычисления параметров [3]. Поэтому важным является применение дифференциального подхода, дающего возможность определить вероятность правильного распознавания отдельно тестируемого образа. Этот подход даёт возможность построить оптимальный вариант модели СР в условиях малых выборок [4]. Математическую модель СР можно представить в виде некоторого функционала

$$M = R(f_x, n, s, t), \tag{1}$$

где f_x — обобщённый классификатор (далее — просто классификатор); n — количество классов; s — размер класса; t — размер доверительного интервала. Модель классификатора f_x в общем виде представляется как

$$f_x = f(\psi^{[k]}, L_{xy}, h_x), \tag{2}$$

где $\psi^{[k]}$ — фрагменты функций признаков; L_{xy} — метрика в пространстве признаков; h_x — решающая функция или правило.

Если для оптимизации модели классификатора использовать последовательный анализ, а в качестве параметра оптимизации — средний размер класса в виде $s = f(\psi^{[k]}, L_{xy}, h_x)$, то задача оптимизации представляется следующим образом:

$$\underset{\psi^{[k]}, L_{xy}, h_x}{\arg\min} \; f(\psi^{[k]}, L_{xy}, h_x) = \min(s). \tag{3}$$

Модель СР включает модель классификатора с параметрами, а также компоненты, влияющие на достоверность результатов распознавания. При оптимизации модели СР важно отдельно учесть влияние на достоверность распознавания таких факторов, как мера расстояний между образами и размеры доверительного интервала и класса, связанные между собой. Основная трудность при исследовании указанных зависимостей состоит в том, что существуют влияния компонент СР как одной на другую, так и совместно — на функционал (1). Всё это усложняет построение моделей различного назначения.

Рассмотрим выражение мер расстояний между векторами признаков $\mathbf{x}$ и $\mathbf{y}$, используемых в теории распознавания [5], через меру Манхеттена — простую линейную меру с весовыми коэффициентами a_i:

$$d(x,y) = \sum_{i=1}^{n} a_i |x_i - y_i|, \tag{4}$$

где $d(x,y)$ — произвольная мера расстояний между векторами $\mathbf{x}$ и $\mathbf{y}$.

Меру расстояний Минковского, как наиболее обобщённую меру, используемую в теории распознавания образов, можно представить в виде

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{n} a_i |x_i - y_i|\right)^{\frac{1}{p}} = C(p) \sum_{i=1}^{n} a_i |x_i - y_i|, \tag{5}$$

где $C(p) = \sum_{i=1}^{n} a_i |x_i - y_i|)^{\frac{1-p}{p}}$; $a_i = |x_i - y_i|)^{p-1}$; $p > 0$.

Из приведенного выше следует, что произвольная метрика — это фильтр в пространстве признаков, т. е. она устанавливает веса признаков при использовании решающих правил. Вес определённого признака должен быть пропорционален приращению одного из показателей при его добавлении в общее признаковое множество, используемое для дискриминации классов: вероятности правильного распознавания, среднего размера класса, дивергенции между классами или дискриминанта Фишера [1, 4, 5]. Можно использовать и другие показатели, однако способ их применения должен быть одинаковым. Если признак не даёт приращения соответствующего показателя или ухудшает его, то значение веса соответствующего признака следует принять равным нулю. Таким образом, путём дополнительного уменьшения количества признаков можно ускорить процесс распознавания, не ухудшая его качественных характеристик. Проблема оптимизации набора признаков и выбора вида метрики решена однозначно с помощью взвешенных признаков и простой линейной меры подобия между образами с весовыми коэффициентами. Задача селекции признаков в этом случае решается частично. Определяется субмножество признаков из генеральной совокупности, выбираемой при помощи того или иного алгоритма (например, ряда ортогональных преобразований). Этот алгоритм в свою очередь должен удовлетворять определённым требованиям относительно селекции признаков — таким, как минимизация энтропии образов класса или максимум дивергенции между классами. Указанным требованиям удовлетворяет метод главных компонент [5].

Последним параметром, используемым в модели, является решающая функция или правило. Условно все решающие функции можно разделить на те, что работают в признаковом пространстве и те, которые

строятся на основании функции расстояний. В признаковом пространстве, например, применяют байесовский классификатор, линейный дискриминант Фишера, метод опорных векторов, и др. В многомерном признаковом пространстве значительно усложняется процедура принятия решения при использовании этих решающих правил. Это особенно нежелательно в случаях, когда распознавание проводится непрерывно для серии образов, поступающих в блок распознавания соответствующей системы. Поэтому при практической реализации СР, работающих с достаточно большиим сериями изображений, используют решающие правила, построенные на основании функции расстояний. Принято использовать два решающих правила: по минимуму расстояния от ближайшего (1NN) и k ближайших соседей (kNN). Хотя 1NN правило наиболее простое, оно характеризуется наименьшими показателями вероятности при принятии решений. Поэтому целесообразно использовать kNN правило. При этом задача сводится к выбору значения k, оптимального для принятия решения в пределах доверительного интервала, соответствующего списку возможных претендентов. От размера класса также зависит размер доверительного интервала, в котором принимается решение. Определение условий, при которых результаты принятия решения на основании оптимального байесовского решающего правила с одной стороны, и 1NN или kNN правил — с другой, совпадают или близки, даёт возможность использовать наиболее простое решающее правило при сохранении качественных свойств процедуры принятия решения.

Литература

[1] *Журавлев Ю. И., Рязанов В. В., Сенько О. В.* Распознавание. Математические методы. Программная система. Практические применения — Москва: Фазис, 2005. — 159 с.

[2] *Капустий Б. Е., Русын Б. П., Таянов В. А.* Новый подход к определению вероятности правильного распознавания объектов множеств // УСиМ. — 2005. — № 2. — С. 8–13.

[3] *Kapustiy B. O., Rusyn B. P., Tayanov V. A.* Tayanov Comparative analysis of different estimates of Recognition Probability // Journal of Automation and Information Sciences. — 2006. — Issue 8. — P. 8–16.

[4] *Kapustiy B. O., Rusyn B. P., Tayanov V. A.* Classifier optimization in small sample size condition // Avtomatika i vychislitel'naya tekhnika. — 2006. — vol. 40, Issue 5. — P.25–32.

[5] *Webb R. A.* Statistical Pattern recognition. — John Wiley & Sons Inc, 2nd ed., 2002.