НЕПАРАМЕТРИЧЕСКИЕ СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ

А.В. Лапко, В.А. Лапко, С.В. Ченцов

Институт вычислительного моделирования СО РАН,
Красноярский государственный технический университет
Chen@fipu.kgtu.runnet.ru

ВВЕДЕНИЕ

Внимание исследователя всегда привлекали методы обработки данных, ориентированные на достаточно низкий уровень априорной информации, что объясняется не только распространенностью в практике подобных условий, но и возможностью построения универсальных алгоритмов, независящих от природы анализируемых объектов. Указанные особенности свойственны непараметрическим моделям и алгоритмам. Их применение не требует введения системы предположений для подгонки объективной реальности под узкие рамки конкретного метода. Основываясь в значительной степени на обучающихся выборках, они позволяют получить результаты максимально адекватные действительности.

НЕПАРАМЕТРИЧЕСКИЕ МОДЕЛИ КОЛЛЕКТИВНОГО ТИПА

Принципы коллективного оценивания нашли широкое распространение на завершающем этапе формирования теории адаптивных систем, когда возникла необходимость обобщения либо получения интегрированных знаний в задачах исследования систем.
В предлагаемом подходе составляющие коллектива представляют собой упрощенные варианты решающих правил, количество которых соизмеримо с объемом обучающей выборки. Следует ожидать, что подобные алгоритмы принятия решений адекватны уровню априорной неопределенности, соответствующему локальным аппроксимациям, и обобщают последние.
Модели восстановления стохастических зависимостей. Пусть задана выборка $V=(x^i, y^i, i=\overline{1,N})$ из статистически независимых наблюдений неизвестной зависимости

$$y=f(x) \quad \forall \quad x \in R^k. \qquad (1)$$

Поставим в соответствие каждой точке обучающей выборки (x^i, y^i) некоторую аппроксимацию $\varphi_i(x, \alpha')$ зависимости (1), параметры которой удовлетворяют условиям

$$y^i = \varphi_i(x^i, \overline{\alpha}^i)$$

$$\overline{\alpha}^i = \mathbf{argmin} \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} (y^j - \varphi_i(x^j, \alpha))^2, i = \overline{1,n}, n << N.$$

Тогда непараметрический коллектив представляется в виде

$$\bar{y} = \bar{f}(x) = \sum_{i=1}^{n} \varphi_i(x, \overline{\alpha}^i) \lambda^i(x),$$

где положительная, ограниченная значением единица, функция $\lambda^i(x)$ определяет "вес" правила $\varphi_i(x, \alpha^i)$ при формировании решения в ситуации x. Примером функции $\lambda^i(x)$ является нормированное расстояние между точками (x, x^i) либо "весовая" функция

$$\lambda^i(x) = \frac{\prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c_v}\right)}{\sum_{j=1}^{n} \prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^j}{c_v}\right)},$$

составленная из “ядерных” функций $c_v^{-1}\Phi\left(\frac{x_v - x_v^i}{c_v}\right)$.

Алгоритмы распознавания образов. Пусть $V = (x^i, \sigma(x^i), i = \overline{1, N})$ обучающая выборка, составленная из параметров складывающейся ситуации x^i и соответствующих им “указаний учителя” $\sigma(x^i)$ о принадлежности i-й ситуации к одному, например, из двух классов.

Следуя методике синтеза коллективов решающих правил, для каждой опорной точки построим линейные решающие функции $\varphi_{12}^i(x, \overline{\alpha}^i)$. Тогда решающее правило, построенное на ее основе, имеет вид

$$m^i(x): \begin{cases} x \in \Omega_1, \text{ если } \varphi_{12}^i(x, \overline{\alpha}^i) \le 0, \\ x \in \Omega_2, \text{ если } \varphi_{12}^i(x, \overline{\alpha}^i) > 0. \end{cases}$$

Параметры i-й решающей функции находятся из условия минимума эмпирической ошибки распознавания образов.

С этих позиций непараметрический коллектив решающих функций в двуальтернативной задаче распознавания образов запишется как

$$\overline{f}_{12}(x) = \left(n\prod_{v=1}^{k} c_v\right)^{-1} \sum_{i=1}^{n} \varphi_{12}^i(x, \overline{\alpha}^i) \prod_{v=1}^{k} \Phi\left(\frac{x_v - x_v^i}{c_v}\right).$$

Отличие от традиционной непараметрической байесовой оценки решающей функции заключается в замене “указаний учителя”

$$\delta(x^i) = \begin{cases} -1, \text{ если } x \in \Omega_1, \\ +1, \text{ если } x \in \Omega_2 \end{cases}$$

на упрощенные решающие функции $\varphi_{12}^i(x, \overline{\alpha}^i)$, $i = \overline{1, n}$.

Обобщенное решение формируется с учетом знака уравнения $\overline{f}_{12}(x)$. Статистика $\overline{f}_{12}(x)$ может быть представлена в виде непараметрической оценки решающей функции и слагаемого, которое стремится к нулю с ростом n. При этом вид упрощенных решающих функций оказывает несущественное влияние на качество распознавания образов.

АНАЛИЗ СЛУЧАЙНЫХ МНОЖЕСТВ

Пусть состояние исследуемого объекта характеризуется случайными множествами $X \subset R^k$ и $Y \subset R^l$, взаимосвязь между которыми определяется неизвестным преобразованием

$$R: X \to Y. \qquad (2)$$

Априорную информацию составляют n пар множеств $(X^i, Y^i, i = \overline{1, n})$, где множеству X^i соответствует вполне определенное множество Y^i.

Традиционные методы построения модели $\overline{R}$ преобразования R основываются на использовании аппарата теории случайных множеств. Идея предлагаемого подхода заключается в замене операций над случайными множествами на менее трудоемкие и хорошо разработанные операции над функциями. Определим на элементах множеств X^i, Y^i непараметрические оценки плотностей вероятности $\overline{p}^i(x)$ $\forall x \in X$ и $\overline{p}^i(y)$ $\forall y \in Y$, $i = \overline{1, n}$. Тогда модель преобразования случайных множеств (2) запишется в виде

$$\overline{\overline{p}}(y) = \frac{\sum_{i=1}^{n} \overline{p}_{ni}(y)\, h(X, X^i)}{\sum_{i=1}^{n} h(X, X^i)},$$

где

$$h(X, X^i) = \begin{cases} \dfrac{8c}{(1-2c)^2}\left(\dfrac{\bar{\rho}^i}{c} - 1\right), & \text{если } \bar{\rho}^i > c \\ 0, & \text{если } \bar{\rho}^i \le c \end{cases}$$

- меры близости между случайными множествами X, X^i, которые формируются по результатам решения задач распознавания образов. Здесь ρ^i - ошибка распознавания в случае двух классов X, X^i, а выбор порогового значения c осуществляется из условия минимума ошибки оценивания $\bar{\bar{p}}(y)$.

МНОГОУРОВНЕВЫЕ СИСТЕМЫ РАСПОЗНАВАНИЯ ОБРАЗОВ

Предлагаются методы синтеза многоуровневых непараметрических систем распознавания образов в условиях больших выборок на основе последовательных процедур формирования решений. По сравнению с прямыми методами обработки информации предлагаемый подход позволяет на порядок повысить вычислительную эффективность систем классификации, создает условия привлечения дополнительных сведений и доверительного оценивания непараметрических решающих правил.

Широкое распространение последовательных методов обработки данных в задачах распознавания образов объясняется возможностью их разбиения на ряд задач принятия решений $m(x) = \{m_t(x(t)), t = \overline{1,T}\}$ по ограниченным наборам признаков сигнала $x = (x(t), t = \overline{1,T})$. Каждый этап процесса обработки сигнала реализуется решающим правилом

$$m_t(x(t)): \begin{cases} x \in \Omega_j, \text{если } \bar{f}_{jm}(x(t)) < 0 \text{ и } p_m(x(t)) = 0, \\ \text{использовать } \ m_{t+1}(x(t+1)), \text{если } x(t) \in \Omega_{jm}(x(t)). \end{cases} \qquad (3)$$

Здесь алгоритм $m_{t+1}(x(t+1))$ осуществляет распознавание в пространстве признаков $x(t+1)$ при условии принадлежности $x(t)$ области пересечения классов $\Omega_{jm}(x(t)) = \Omega_j \cap \Omega_m$. Непараметрическая оценка решающей функции определяется статистикой

$$\bar{f}_{jm}(x(t)) = \left[I_{jm}(t)\right]^{-1} \sum_{i \in I_{jm}(x)} \sigma(x^i(t))\, \Phi\left(\frac{x(t) - x^i(t)}{c}\right),$$

где $\Phi(\cdot)$ – многомерная функция ядерного типа удовлетворяющая условиям положительности, симметрии и нормированности;

$[I_{jm}]$ – количество точек исходной обучающей выборки из области $\Omega_{jm}(x(t))$;

$$\sigma(x_i(t)) = \begin{cases} -1 & \forall\, x_i \in \Omega_j \\ 1 & \forall\, x_i \in \Omega_m \end{cases}.$$

Использование иерархических структур в процессах распознавания создает предпосылки рационального использования дополнительной информации о ранее вскрытых закономерностях в пространстве признаков сигнала $x(t)$, $t = \overline{1,T}$ и существенно снизить время q_n решения задачи распознавания по сравнению со временем q прямой обработки сигнала x.

$$q_n / q = \sum_{t=1}^{T} K(t) P\{x(t) \in \Omega_{jm}(\bar{x}(t))\} / K < 1,$$

$$x(t) = (x_v(t),\ v = \overline{1, K(t)}),\ K = \sum_{t=1}^{T} K(t)$$

и при равных размерностях $K(t)=K/T$ наборов признаков $x(t)$, $t = \overline{1,T}$ сигнала x не превышает величину

$$(1 - P^T(1))/[T(1 - P(1))],\ P(1) = P\{x(1) \in \Omega_{jm}(x(1))\}.$$

Выбор рациональной структуры систем. Если информация о классифицируемом объекте поступает последовательно в дискретном времени, то число уровней структуры определяется количеством этапов поступления новых данных. В других случаях формирование структуры многоуровневого алгоритма распознавания образов и распределение признаков по конкретным этапам обработки информации является непростой задачей. Предлагается и исследуется оригинальный метод минимизации описания исходного пространства признаков, основанный на принципах обучения и аппарате теории графов.
Для создания оптимальной по времени структуры системы распознавания образов необходимо распределить на первых уровнях структуры наиболее информативные наборы признаков.
Модель формирования наборов статистически независимых признаков:
- вычислить ошибку распознавания образов ρ_j, $j = \overline{1,k}$ в пространстве каждого признака x_j, $j = \overline{1,k}$;
- сформировать всевозможные пары признаков x_i, x_j и вычислить ошибку распознавания в их пространстве $\rho_{ij}, i, j = \overline{1,k}$, $i \neq j$;
- провести анализ полученных результатов. Построить граф, в котором вершины соответствуют исходным признакам. Между двумя вершинами графа x_i, x_j имеется ребро, если произведение ρ_i ρ_j ошибок распознавания в пространстве соответствующих признаков достоверно не отличается от ошибки распознавания ρ_{ij} в пространстве двух этих признаков с некоторым уровнем доверия β;
- решить задачу декомпозиции графа на компоненты, обладающие свойством сильной связности. В таких подграфах каждая вершина соединена ребрами со всеми остальными вершинами.
Наборы признаков, соответствующие выделенным компонентам определяют количество и распределение признаков между уровнями структуры системы классификации.

ЗАКЛЮЧЕНИЕ

Рассмотренные направления обработки информации объединяют две основные идеи: создание условий для применения традиционных непараметрических методов статистики; проведение предварительной обработки информации с целью обнаружения дополнительных сведений, повышающих эффективность решения поставленных задач. Выбор того или иного метода зависит от конкретного приложения и особенностей априорной информации. При этом учитывая системный характер проблем обработки информации, не исключается использование одновременно нескольких подходов для их преодоления.
Перспективным является развитие исследований по следующим направлениям: синтез и анализ методов на основе последовательных процедур принятия решений; разработка непараметрических моделей коллективного типа при восстановлении многомерных стохастических зависимостей и их систем. Успешное продвижение в данных направлениях позволит создать теоретическую основу исследования сложных систем.

ЛИТЕРАТУРА

1. Лапко А.В., Ченцов С.В., Крохов С.И., Фельдман Л.А. Обучающиеся системы обработки информации и принятия решений (непараметрический подход). Новосибирск: Наука, 1996. 296 с.
2. Лапко А.В. Непараметрические методы классификации и их применение. Новосибирск: Наука, 1993. 152 с.
3. Лапко А.В., Ченцов С.В. Многоуровневые непараметрические системы принятия решений. Новосибирск: Наука, 1997. 250 с.