УДК 519.246
ББК 22.172

Т.И. Ведерникова

СПОСОБЫ ПОСТРОЕНИЯ РАБОЧЕГО СЛОВАРЯ ПРИЗНАКОВ ДЛЯ РЕШЕНИЯ ЗАДАЧ ИДЕНТИФИКАЦИИ

Содержит обзорную информацию о методах и алгоритмах выделения информативных признаков для идентификации стохастических объектов.

Ключевые слова: информативный признак, рабочий словарь признаков, идентификация, векторная случайная величина.

T.I. Vedernikova

WAYS OF CREATING A WORKING DICTIONARY OF FEATURES FOR SOLVING IDENTIFICATION PROBLEMS

The article provides general overview information about the methods and algorithms of determining informative features for identification of stochastic objects.

Keywords: informative feature, working dictionary of features, identification, vector random variable.

Одной из проблем, возникающих при решении задач классификации и распознавания образов, является выбор наиболее подходящей математической модели обработки и описания данных. При этом соответствующая информация представляет собой многомерные временные ряды, и, довольно часто, в общем потоке имеющихся данных исследователя интересует лишь та часть данных, которая предшествует аномальному явлению и следует после него. Идентификация аномального явления по совокупности наблюдений представляет собой задачу распознавания образов, эффективность решения которой существенно зависит от качества рабочего словаря признаков (РСП).

Если имеется достаточно полный словарь информативных признаков, то собственно распознавание и идентификация уже не вызывают особых затруднений. В большинстве практических задач распознавания определение полного набора различных признаков, описывающих объект, оказывается делом исключительно трудным или практически невозможным. С другой стороны, большой объем наблюдаемых параметров может «зашумлять» картину идентификации объектов и увеличивать время на выполнение самой процедуры распознавания.

Для выполнения дальнейших рассуждений дадим ряд определений.

Вектор $X = (x_1, x_2, \ldots, x_n)$, полученный в результате непосредственных измерений (статистические данные), — есть вектор параметров объекта.

Вектор признаков — есть новый вектор $Y = (y_1, y_2, \ldots, y_m)$, представляющий собой либо усеченный набор параметров, либо вектор, получаемый путем некоторого преобразования параметров $y_i = \phi_i(X)$.

Априорный словарь признаков (АСП) — это совокупность всевозможных параметров и признаков, относящихся к объекту.

Рабочий словарь признаков — это совокупность наиболее информативных параметров и/или признаков (часть АСП), сформированная на этапе обучения с целью решения задачи идентификации. Построить

ЭЛЕКТРОННЫЙ НАУЧНЫЙ ЖУРНАЛ
ИЗВЕСТИЯ ИРКУТСКОЙ ГОСУДАРСТВЕННОЙ ЭКОНОМИЧЕСКОЙ АКАДЕМИИ (Байкальский государственный университет экономики и права)
2012. № 4
http://eizvestia.isea.ru

рабочий словарь признаков — значит найти такую систему признаков, которая достаточно полно описывает объект с точки зрения заданного критерия информативности.

Эффективность признака определяется величиной полезного вклада при распознавании. Оценкой полезного вклада может служить: дивергенция, энтропия, прямая оценка вероятности ошибки. В том случае, когда распределения вероятностей признаков известны, пользуются критериями дивергенции и энтропии. Если распределения вероятностей признаков для каждого класса не известны, то используют критерии, основанные на прямой оценке вероятности ошибки идентификации, где в зависимости от выбранного способа преобразования пространства параметров X в пространство признаков Y отыскивается экстремальное значение заданного критерия эффективности признаков.

Существующие эвристические алгоритмы и методы построения РСП для описания образов можно разделить на две большие группы: 1) методы минимизации систем описания образов (объектов исследования), когда РСП есть подпространство Y (размерности m) полного признакового пространства X экономического объекта (размерности n), $m < n$; 2) структурные методы описания образов, когда признаковое пространство Y получается путем некоторого преобразования априорного пространства X.

Методы минимизации систем описания образов. Эта группа методов включает: алгоритмы направленного перебора различных подсистем параметров с целью выбора наилучшей с точки зрения заданного критерия; алгоритмы выбора информативных признаков на основе расстояния между параметрами различных классов; алгоритм случайного поиска с адаптацией (игровой алгоритм).

Под минимизацией описания обычно понимается уменьшение числа измеряемых параметров. При этом либо заранее (из каких-либо физических соображений) определяется количество признаков $m (m < n)$, либо m выбирают в зависимости от требуемой точности решения задачи идентификации. Если значения n и m достаточно малы, то выбор системы признаков Y из пространства параметров X можно осуществить путем полного перебора всех возможных комбинаций, равных числу сочетаний из n по m. Для больших значений n и m применяются алгоритмы, не использующие полный перебор. Например, алгоритм Мерилла и Грина (алгоритм последовательного исключения), суть которого заключается в поочередном исключении одного параметра, в отсутствие которого критерий принимает оптимальное значение. Другой алгоритм (алгоритм последовательного включения параметров), не использующий полный перебор, состоит в том, что система признаков формируется по принципу включения в нее наилучшего по заданному критерию информативности параметра.

«Расстояния» играют важную роль при обработке информации. Алгоритмы выбора информативных признаков основываются на расстояниях между параметрами внутри классов (однородных объектов), так и на расстояниях параметров между классами. При этом в качестве информативных признаков выбираются те параметры, расстояния между которыми внутри класса минимальны, а между классами максимальны.

В основу алгоритма случайного поиска с адаптацией положено предположение о том, что признаки, входящие в наиболее информативную систему, чаще встречаются в тех системах, которые близки к ней по некоторому критерию информативности и наоборот.

ЭЛЕКТРОННЫЙ НАУЧНЫЙ ЖУРНАЛ
ИЗВЕСТИЯ ИРКУТСКОЙ ГОСУДАРСТВЕННОЙ ЭКОНОМИЧЕСКОЙ АКАДЕМИИ (Байкальский государственный университет экономики и права)
2012. № 4
http://eizvestia.isea.ru

ЭЛЕКТРОННЫЙ НАУЧНЫЙ ЖУРНАЛ
ИЗВЕСТИЯ ИРКУТСКОЙ ГОСУДАРСТВЕННОЙ ЭКОНОМИЧЕСКОЙ АКАДЕМИИ (Байкальский государственный университет экономики и права)
2012. № 4
http://eizvestia.isea.ru

Структурные методы описания образов. Особенность структурных методов описания образов состоит в том, что РСП включает не непосредственные измерения (параметры), а признаки, отражающие типологию объектов, принадлежащих одному классу. К структурным методам относятся: методы факторного анализа; методы нелинейного преобразования пространства параметров (многомерное масштабирование); методы аппроксимации параметров; методы вейвлет-анализа; МОС-метод (метод моделирования статистик, различающих случайные величины).

Основная идея факторного анализа состоит в том, чтобы, наблюдая большое число измеряемых параметров, выявить меньшее число таких признаков (факторов), которые в основном определяют поведение параметров и характеризуют исследуемый объект.

Процесс нелинейного преобразования пространства параметров X в пространство признаков Y (меньшей размерности) называют еще многомерным масштабированием. Целью нелинейных отображений является изменение структуры расстояний между объектами с соблюдением ограничений, накладываемых структурой исходного распределения, и предоставление возможности наглядного анализа многомерных данных. Алгоритмы нелинейного преобразования параметров можно разделить на две группы: первые осуществляют преобразование $y_i = \phi_i(X)$ с одновременным понижением размерности признакового пространства (итеративный и неитеративный алгоритмы), вторые производят двумерное отображение.

К методу стохастической аппроксимации прибегают в том случае, когда наблюдаемые в выбранных точках X_j значения функции параметров $f_j = f_j(X)$ являются реализациями случайных величин ξ. В качестве критерия эффективности берутся математические ожидания $M_i[\xi]$, $i = \overline{1,\ M}$. Задача выделения признаков сводится к отысканию наилучшей аппроксимирующей функции $\widehat{f}_i = \widehat{f}_i(X)$, минимизирующей критерий.

Вейвлет-анализ применяется для изучения структуры неоднородных процессов. Вейвлет-преобразование заключается в разложении одномерной функции по базису, сконструированному из обладающей определенными свойствами солитоноподобной функции (вейвлета) посредством масштабных изменений и переносов.

Метод моделирования статистик, различающих случайные величины, базируется на переходе от многомерных наборов признаков, описывающих объекты,

$$\overline{X_k} = (x_{k1}, x_{k2}, \ldots, x_{kn}) \qquad (1)$$

к одномерным $\{Y_{ik}\}$; $i = \overline{1,\ m}$; $k = \overline{1,\ M}$, аккумулирующим отличительные особенности наблюдений (1). Основополагающим для метода является предположение о том, что разные объекты имеют разные наборы признаков, т.е. являются реализациями разных случайных величин. Следовательно, для каждой совокупности однородных объектов существуют свои характерные преобразования вида

$$Y_{ik} = \phi_i(X_k) = (x_{k1},\ x_{k2},\ \ldots,\ x_{kn}),\ \ i = \overline{1, m};\ \ k = \overline{1, M}, \qquad (2)$$

отражающие внутреннюю структуру объекта идентификации. РСП одинаковых объектов составляет характеристика или небольшой набор характеристик, определенных экспертным путем или выявленных в результате обучения.

В основу различения признаков типа (1) положены отношения $Z_o = \dfrac{a}{b}$ и модули разности $Z_p = |a - b|$, где в качестве a и b берутся состав-

ляющие вектора (1). Базовыми принципами МОС-метода являются: одномерность признаков РСП; избыточность признаков, описывающих объект; минимальная вариабельность признака; «непересекаемость» распределений признаков, аккумулирующих отличительные особенности разных объектов.

Реализация этих принципов осуществляется следующим образом. Каждый признак формируется как функция типа (2). Уход от многомерности позволяет упростить процесс решения задачи идентификации, так как анализ многомерных наблюдений выполнить много сложнее, чем одномерных. Многообразие признаков определяется интуицией и опытом исследователя, а избыточность признаков легко достигается, так как количество функций вида (2) ничем не ограничено. В основу принципа «непересекаемости» положена идея о том, что всегда можно подобрать такие преобразования $\phi_i(\bullet)$; $i = 1, m$, при которых получаемые признаки будут иметь минимальную вариацию, а для разных объектов еще и значительные количественные различия.

Проведенный обзор методов выделения набора информативных признаков (построения РСП) для решения задач идентификации позволяет сделать следующие выводы.

Существенным недостатком методов минимизации систем описания образов является то, что они не отражают внутреннюю структуру объекта идентификации. Кроме того, алгоритмы перебора требуют, как правило, много времени. Алгоритмы «расстояний» обычно узконаправленны, т.е. хорошо решают конкретную задачу идентификации, а для другого приложения могут быть совершенно неэффективными. РСП, построенный посредством игрового алгоритма, не всегда соответствует реальной действительности.

Классические методы факторного анализа предназначены для решения статических задач, хотя иногда их используют для анализа «срезов» временных точек. Что касается итеративного алгоритма, то в случае больших последовательностей неклассифицированных данных он может оказаться «вычислительным монстром», то же касается и вейвлет-анализа. В неитегративном алгоритме есть существенное ограничение — это предположение о том, что опорные точки существуют, и расстояния между признаками и опорными точками заданы. Большим недостатком алгоритмов аппроксимации является допущение о наличии аппроксимирующей функции, которой, однако, может и не быть.

Сказанное выше позволяет предложить МОС-метод как эффективный инструмент для решения задач идентификации многомерных стохастических объектов.

Информация об авторе

Ведерникова Татьяна Ивановна — кандидат технических наук, доцент, кафедра информатики и кибернетики, Байкальский государственный университет экономики и права, г. Иркутск, e-mail: vedernikova-ti@isea.ru.

Author

Vedernikova Tatiana Ivanovna — PhD in Engineering Science, Associate Professor, Chair of Computer Science and Cybernetics, Baikal State University of Economics and Law, Irkutsk, e-mail: vedernikova-ti@isea.ru.

ЭЛЕКТРОННЫЙ НАУЧНЫЙ ЖУРНАЛ
ИЗВЕСТИЯ ИРКУТСКОЙ ГОСУДАРСТВЕННОЙ ЭКОНОМИЧЕСКОЙ АКАДЕМИИ (Байкальский государственный университет экономики и права)
2012. № 4
http://eizvestia.isea.ru