ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 22 МАТЕМАТИКИ Выпуск 4
2015

Т. В. Жгун, А. В. Липатов (Великий Новгород, НовГУ)**. Определение информативности при вычислении интегральной характеристики изменения качества системы при решении задачи выделения сигнала в условиях априорной неопределенности.**

Одной из центральных проблем при решении задач обработки информации является проблема выбора информативного подмножества признаков и оценки его пригодности. Одним из широко распространенных методов сокращения размерности изображений является метод главных компонент. Распространенный способ выбора числа главных компонент — оставить число главных компонент, которые объясняют заданный процент общей дисперсии [1]:

$$\gamma_\sigma = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_l}{\lambda_1 + \lambda_2 + \cdots + \lambda_n} \geqslant \theta. \tag{1}$$

Однако направления, максимизирующие дисперсию, далеко не всегда максимизируют информативность. Может случиться, что именно младшие главные компоненты несут необходимую смысловую нагрузку. На странице сайта Alglib [2] приводится пример, когда главная компонента с максимальной дисперсией не несет почти никакой информации, в то время как компонента с минимальной дисперсией позволяет полностью разделить классы.

Подходы к оценке числа главных компонент по необходимой доле объясненной дисперсии формально применимы всегда, однако неявно они предполагают, что нет разделения на «сигнал» и «шум», и любая заранее заданная точность имеет смысл. При разделении данных на полезный сигнал и шум задаваемая точность теряет смысл, и требуется переопределить понятие информативности. Считая, что определяемая интегральная характеристика есть слабый полезный сигнал, который нужно распознать в зашумленных данных (например, в статистических), получаем решение при помощи ОСШ-алгоритма. Интегральная оценка системы из m объектов, каждый из которых характеризуется n признаками, для момента t имеет вид [3]:

$$-q^t = A^t\, W^*, \tag{2}$$

где $q^t = (q_1^t, q_2^t, \ldots, q_m^t)^T$ — вектор интегральных индикаторов момента t, A^t — матрица предобработанных данных для момента t, $W^* = (w_1^*, w_2^*, \ldots, w_n^*)$ — веса показателей.

При применении дисперсионного критерия информативности выбор порогового значения ОСШ определяет выбор параметра информативности θ, определяющего относительную долю разброса γ, приходящуюся на первые главные компоненты (1). Если информативность γ выражена в долях единицы, то величину отношения сигнал/шум можно представить как отношение полезной части используемой информации γ к неиспользуемой информации $1-\gamma$, т. е. $SNR = \gamma/(1-\gamma)$. Если рассматриваемое значение отношения сигнал/шум не менее порогового значения Θ, то справедлива

оценка информативности

$$\gamma \geqslant \frac{\theta}{\theta + 1}. \tag{3}$$

Соотношение (3) дает априорную оценку снизу информативности выбранной системы признаков в зависимости от используемого значения ОСШ. В табл. 1 представлены некоторые значения, связывающие рассматриваемые показатели.

Таблица 1. Связь дисперсионной информативности с используемым значением ОСШ

SNR	1,2	1,5	2,0	2,2	3,0	4,0	5,7	9,0	19,0
Информативность, γ	0,550	0,600	0,667	0,688	0,750	0,800	0,851	0,900	0,950

Однако величина $SNR = 2,2$ является оптимистичной величиной для статистических данных, и при увеличении этого значения хотя бы до трех единиц большая часть эмпирических главных компонент (ЭГК) окажутся просто нулевыми, что не увеличит их информативности. Очевидно, что ОСШ-информативность ЭГК определяется параметрами найденных действующих переменных. Аналогично дисперсионному критерию информативности (1), можно определить ОСШ-критерий информативности для выбранного числа эмпирических главных компонент:

$$\gamma_{SNR} = \frac{S_{11} + S_{12} + \cdots + S_{1N}}{S_{21} + S_{22} + \cdots + S_{2N}},$$

где $S_{1k}\ \ldots$ — сумма величин ОСШ у действующих переменных k-й ЭГК, S_{2k} — сумма ОСШ всех переменных k-й ЭГК. Тогда информативность выбранной системы признаков определяется дисперсионной и ОСШ-информативностью: $\gamma = \gamma_\sigma\, \gamma_{SNR}$.

Таблица 2. Определение информативности интегрального показателя «Качество населения»

	Номер эмпирической главной компоненты								
	1	2	3	4	5	6	7	8	9
Сумма ОСШ эмпирической ГК	99,2	67,1	30,3	22,8	12,9	12,9	21,9	46,7	59,7
Сумма действующих ОСШ	98,6	64,5	23,6	20,0	10,1	8,1	17,0	43,2	55,6
Накопленные %%, γ_{SNR}	26,4	43,7	50,0	55,3	58,0	60,2	64,8	76,3	91,2
Эмпирические собственные числа	3,1	2,0	1,1	0,8	0,6	0,5	0,5	0,3	0,1
Накопленные %%, γ_σ	34,1	56,6	69,2	78,0	85,0	90,7	95,8	99,1	100,0
Информативность, γ, %	9,0	24,7	34,6	43,1	49,3	54,6	62,0	75,6	91,2

В табл. 2 приведен пример определения информативности интегрального показателя «Качество населения», вычисленный по (2). При рассмотрении всех 9 ЭГК суммарная информативность составит около 91%. Число выбираемых ЭГК, определяемых величиной $SNR = 2,2$, согласно (3), должно обеспечивать ОСШ-информативность не менее 68,8%. В данном случае это 8 ЭГК.

СПИСОК ЛИТЕРАТУРЫ

1. *Айвазян С. А.* Интегральные индикаторы качества жизни населения: их построение и использование в социально-экономическом управлении межрегиональных сопоставлениях. М.: ЦЭМИ РАН, 2000, 56 с.
2. Линейный дискриминантный анализ. Alglib. Open sourse. [Электронный ресурс]. Режим доступа: http://alglib.sources.ru/dataanalysis/lineardiscriminantanalysis.php Загл. с экрана (дата обращения: 13.01.2015).

3. *Жгун Т. В.* Построение интегральной характеристики изменения качества системы на основании статистических данных как решение задачи выделения сигнала в условиях априорной неопределенности. — Вестник Новгородского гос. ун-та. Сер.: Технические науки, 2014, № 81, с. 10–16.