ФОРМИРОВАНИЕ ПРИЗНАКОВОГО ПРОСТРАНСТВА ПО КРИТЕРИЮ СОПРЯЖЕННОСТИ ВЕКТОРОВ ИЗМЕРЕНИЙ

В.А. Фурсов, В.А. Шустов
Институт систем обработки изображений РАН

1. Постановка задачи

Проблема формального анализа информативности выбранной системы признаков и обоснования размерности признакового пространства является одной из основных в общей проблеме обучения распознаванию образов. В работе [1] для обучения распознаванию цифр используются меры, основанные на вычислении энтропии. Применение этих мер требует для выявления закономерностей обработки большого числа (ансамбля) реализаций. В работе [2] для выбора размерности признакового пространства при малом числе обучающих объектов предложено использовать меры ориентации векторов признаков относительно нуль-пространства транспонированной матрицы признаков. Эти меры, по существу характеризуют мультиколлинеарность векторов-столбцов, из которых составлена матрица признаков.

В настоящей работе задача формального анализа признакового пространства в задаче, рассматривавшейся в [1] осуществляется с использованием аппарата анализа, предложенного в [2]. При этом по сравнению с [1] изменена также область локализации цифр. В частности, при выборе этой области учитывалось, что вертикальный размер цифр, обычно, больше горизонтального.

2. Основные соотношения

В настоящей работе используется техника вычисления и выбора пространства признаков, основанная на анализе степени сопряженности (мультиколлинеарности) образующих его векторов. Основные соотношения запишем для случая задачи распознавания двух классов. Обобщение на случай многих классов осуществляется как обычно [4].

Задача решается в классе линейных разделяющих функций вида

$$g(\mathrm{x}) = \mathbf{a}^{\mathrm{T}}\mathbf{y}(\mathbf{x}) = \mathrm{w}_0 + \sum \mathrm{w}_\mathrm{i}\mathrm{x}_\mathrm{i},\ \mathrm{i} = \overline{1,\mathrm{d}} \qquad (1)$$

где $\mathbf{a}=[\mathrm{w}_0,\ \mathbf{w}]^{\mathrm{T}}=[\mathrm{w}_0,\mathrm{w}_1,...\mathrm{w}_\mathrm{d}]$ - M-мерный вектор искомых параметров (весовых коэффициентов [4]), а M×1-вектор $\mathbf{y}(\mathbf{x})$ определяется как $[1,\mathbf{x}]^{\mathrm{T}}$. После определения признаков N объектов для обучения системы (N>M) имеем N×M-матрицу **Y**, строками которой являются векторы $\mathbf{y}_i^T, i = \overline{1,N}$ Столбцы матрицы **Y** далее будем обозначать $\mathbf{Y}_j$. $j = \overline{1,M}$.

Задача оценивания параметров классификатора решается следующим образом. Ищется разделяющий вектор **a**, удовлетворяющий (в смысле критерия среднеквадратической ошибки) уравнению

$$\mathbf{Ya}=\mathbf{b}, \qquad (2)$$

где **b** - вектор, формируемый по заданным допускам, определяющим область решений [4]. Предполагается, что число строк N матрицы **Y** не намного превышает число столбцов (признаков) M.

При выборе признакового пространства в рамках указанной постановки задачи необходимо иметь в виду следующую трудность. Состав и количество признаков могут оказаться недостаточно информативными на этапе классификации, в частности, конкретная реализация измерений на системе обучающих объектов может оказаться такой, что соответствующая задача оценивания разделяющей поверхности может оказаться плохо обусловленной или почти вырожденной. Связано это с тем, что векторы $\mathbf{Y}_j$ матрицы **Y** могут оказаться линейно-зависимыми (или "почти" линейно-зависимыми), так что задача оценивания параметров разделяющей функции не может быть решена. Эта проблема известна как проблема анализа мультиколлинеарности системы векторов [3].

Для преодоления этих трудностей в настоящей работе рассматривается возможность использования предложенного в [2] метода формирования пространства признаков, основанного на вычислении мер взаимной ориентации векторов признаков. В частности, ставится задача исследовать возможность применения для определения информативности признаков изображения так называемого показателя сопряженности векторов $\mathbf{Y_i}$ с нуль-пространством матрицы $\mathbf{Y}^{\mathrm{T}}$ [5].

В соответствии с этим методом для каждого вектора $\mathbf{Y_i}$, $\mathrm{i} = \overline{1,\mathrm{M}}$ вычисляется величина

$$\mathrm{S}_\mathrm{i} = \left(\mathbf{Y}_\mathrm{i}\mathbf{T}_{0,\mathrm{M}-1}\mathbf{T}_{0,\mathrm{M}-1}^{\mathrm{T}}\mathbf{Y}_\mathrm{i}^{\mathrm{T}}\right)^{1/2} \Big/ \left(\mathbf{Y}_\mathrm{i}\mathbf{Y}_\mathrm{i}^{\mathrm{T}}\right)^{1/2}, \qquad (3)$$

где $\mathbf{T}_{0,\mathrm{M-1}}$ - N×(N-M+1) - матрица, составленная из (N-M+1) собственных векторов, соответствующих нулевым собственным значениям матрицы $\mathbf{Y}_{\mathrm{M}-1}\mathbf{Y}_{\mathrm{M}-1}^{\mathrm{T}}$. Матрица $\mathbf{Y}_{\mathrm{M}-1}$ получается из матрицы **Y** путем вычеркивания одного, в данном случае, i-го вектора-столбца $\mathbf{Y}_\mathrm{i}$.

С использованием полученных величин S_i далее выбираются состав признаков и размерность признакового пространства. Для этого вычисленные значения S_i сравниваются с заданным допустимым значением $\mathrm{S}_{\text{доп}}$. Если оказывается, что

$$\mathrm{S}_\mathrm{i} < \mathrm{S}_{\text{доп}}, \qquad (4)$$

то столбец $\mathbf{Y}_\mathrm{i}$, исключается. Величина $\mathrm{S}_{\text{доп}}$ определяется экспериментально и не должна быть слишком малой, т.к. малые значения S_i говорят о сильной мультиколлинеарности векторов матрицы Y и, сле-

довательно, о плохой обусловленности задачи оценивания. Основная цель этой работы заключается в установлении порогового значения этого показателя для отбора векторов матрицы **Y** в заданной системе признаков.

3. Описание исходной системы признаков

Рассматривается задача распознавания цифр, записанных с использованием различных шрифтов. Следуя работе [1] система признаков формируется следующим образом. Поле цифры разбивается на квадраты небольших размеров. Фиксируется суммарная величина яркости, приходящаяся на все пикселы, внутри каждого квадрата. Для обозначения квадратиков далее используются две цифры, первая из которых обозначает номер строки (начиная сверху) на растровом изображении, а вторая - номер столбца. Матрица признаков **Y** содержит M столбцов, а число строк - N соответствует числу цифр, предъявляемых для обучения.

Для выбранной системы признаков существенное значение имеют размеры и число квадратов на поле цифры. При разбиении на большое число квадратом малых размеров, наверняка, окажется, что многие из них на поле области, ограничивающей всю совокупность цифр, окажутся малоинформативными. Ясно, что на этапе обучения исключение соответствующих признаков может быть произведено без ущерба для точности классификации, приводя при этом к существенному сокращению размерности пространства решаемой задачи. Более того, это приведет к повышению надежности и быстродействия построенной на выбранной системе признаков процедуры классификации.

Преимущества используемого в настоящей работе подхода к решению задачи формирования системы признаков особенно проявляются в тех случаях, когда обучение распознаванию образов должно осуществляться по малому числу наблюдений. Это может быть связано с нестационарностью распределений образов в признаковом пространстве, требующей частой перенастройки классификатора или с невозможностью получить достаточно большой объем наблюдений. В указанной ситуации использование априорных вероятностных характеристик обучающей выборки для определения информативности признаков [4] не вполне правомерно.

Далее приводятся результаты экспериментов, показывающие возможность снижения размерности указанной системы признаков без ущерба для качества распознавания классов.

4. Результаты экспериментов

Чтобы оценить эффект от использования показателя сопряженности для оценки информативности был проведен эксперимент по классификации цифр. Различные образцы изображения цифр были преобразованы в бинарный растр 32×64. Растр был поделен на фрагменты 8×8 и в качестве признаков изображения использовалось количество черных пикселов в фрагменте. Образцов каждой цифры было 20.

На рисунке 1 приведено поле показателя сопряженности признаков (фрагментов растра) для цифр “2” и “3” (более темный цвет соответствует меньшему значению S_i). Для двух фрагментов все компоненты вектора $\mathbf{Y}_i$ оказались равными нулю (на рисунке соответствующие квадратики отмечены крестиками). Эти признаки были исключены из дальнейшего рассмотрения. Таким образом, матрица **Y** из уравнения 2 имела размерность 40×31.

В таблице приведены минимальные и максимальные значения S_i для матрицы **Y** и **Y**’, полученной после исключения из **Y** восьми столбцов с минимальными значениями S_i ($S_{доп}$ = 0,05). При этом условию (4) удовлетворяли показатели сопряженности столбцов, соответствующих на рисунке фрагментам 13, 24, 34, 64, 74, 81, 82 и 83. Как видно, после исключения группы признаков их сопряженность значительно уменьшилась.

Рис. 1. Поле показателей S_i

Предельные значения показателей S_i

	Y	**Y**’
S_{min}	0,026	0,092
S_{max}	0,210	0,510

Проводилось также сопоставление количества исключенных по показателю сопряженности признаков с количеством неправильных классификаций. Оказалось, что все объекты классифицируются правильно при исключении не более 10 признаков с наименьшими показателями сопряженности. Заметное число ошибок классификации стало появляться при исключении более 15 признаков.

Работа выполнена при поддержке РФФИ, грант 99-01-00079.

Заключение

Методика формирования системы признаков с использованием мер сопряженности векторов с нуль-пространством позволяет осуществлять обоснованный выбор числа признаков, в частности, снижение размерности задачи без ущерба для качества классификации. В принятой системе признаков может также решаться задача выбора количества фрагментов, на которые разбивается поле цифры.

Литература

1. Chi Z., Yan H., Feature evaluation and selection based on an entropy measure with data clustering// Optical Engineering-1995, Vol. 34 No. 12, p. 3514-3519.
2. Фурсов В.А. Метод проекций на нуль-пространство в проблеме распознавания образов по малому числу наблюдений. Труды Всероссийской конференции “Математические методы распознавания образов” (ММРО-9), Москва, 15-19 ноября, 1999 г., с. 119-121.
3. Демиденко Е.З. Линейная и нелинейная регрессии. М.: Финансы и статистика, 1978 г.
4. Р. Дуда, П. Харт. Распознавание образов и анализ сцен. Пер с англ. М.: Мир, 1976, 512 с.
5. Фурсов В.А. Идентификация моделей систем формирования изображений по малому числу наблюдений.– Самара: Издательство Самарского государственного аэрокосмического университета им. Академика С.П. Королева, 1998.– 218 с.