

ОТБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ ПРИ ОБУЧЕНИИ РАСПОЗНАВАНИЮ ОБРАЗОВ С УПОРЯДОЧЕННЫМИ ПРИЗНАКАМИ

© Середин О.С.¹, Моттль В.В.²

¹ Тульский государственный университет,
Россия 300600, г.Тула, пр.Ленина, 92

² Вычислительный центр РАН,
Россия 117967, г.Москва, ГСП-1, ул.Вавилова, 40

E-MAIL: oseredin@yandex.ru, vmottl@yandex.ru

Abstract. The task of interrelated features selection is considered. A lot of approaches assume that feature vector is unordered set of numerical coefficients. However in some tasks features are serial measurement along some axes, for example, counts of some kind of signal. The technique of feature selection using a priori information about one-dimensional order is suggested.

ВВЕДЕНИЕ

Проблема решения задачи обучения распознаванию образов в случае большого числа признаков (по сравнению с количеством объектов в обучающей выборке, так называемая ситуация «проклятия размерности»), как правило, сводится к паре стандартных подходов – сокращение размерности за счет отбора наиболее информативных признаков или наложение на решающее правило априорных ограничений (регуляризация решающего правила распознавания). В статье предложен подход, комбинирующий две эти методики. Рассматривается отбор информативных признаков с учетом их упорядоченности в задаче обучения распознаванию образов. Большинство подходов к отбору информативных признаков рассматривает вектор признаков объектов, как неупорядоченную совокупность числовых коэффициентов, более того, многие методы при своей постановке явно принимают гипотезу о том, что отдельные компоненты вектора признаков являются независимыми. Однако в некоторых задачах признаки – есть суть последовательных измерений вдоль оси некоторого аргумента, например, упорядоченные отсчеты какого-либо сигнала, компоненты спектра и т.п.

В ряде предыдущих статей авторов были предложены как способы регуляризации решающих правил распознавания, так и способы отбора признаков. Так в ранних работах [2, 3] был предложен способ регуляризации решающего правила распознавания, учитывающий априорную информацию о взаимосвязи признаков. Одновременно проводились исследования по способам комбинирования модальностей в задачах интеллектуального анализа данных, которые фактически предложили эффективный инструмент отбора информативных признаков [4, 5, 6]. В этой работе предлагается осуществлять отбор признаков с учетом априорной информации об одномерной упорядоченности признаков вдоль оси. Следует отметить, что в качестве теоретической базы предлагаемых алгоритмов выбран метод потенциальных функций, и вытекающий из него метод опорных векторов [1], как получивший в последнее время распространение и популярность.

Статья построена следующим образом – в первом разделе обосновано использование в качестве базы модельных модификаций популярного метода опорных векторов, в следующих двух разделах отдельно рассматриваются алгоритмы отбора признаков и регуляризации решающего правила распознавания, четвертый посвящен описанию нового правила отбора признаков в случае их упорядоченности в задаче обучения распознаванию образов. В заключении обсуждаются перспективы развития предложенной методики.

1. МЕТОД ОПОРНЫХ ВЕКТОРОВ – ОСНОВА МОДИФИКАЦИЙ

Пусть имеется обучающая выборка $\{\mathbf{x}_j, g_j\}, j = 1, \dots, N$, где $\mathbf{x} = (x_i, i = 1, \dots, n) = \mathbf{x}(\omega) \in R^n$ – исходный действительнзначный вектор признаков объекта распознавания ω , $g = \{\pm 1\}$ – индекс классификации каждого объекта. Направляющий вектор оптимальной разделяющей гиперплоскости $\mathbf{a} = \{a_1, a_2, \dots, a_n\}$ и её смещение b определяются как решение хорошо известного критерия:

$$\begin{cases} \mathbf{a}^T \mathbf{a} + C \sum_{j=1}^N \delta_j \rightarrow \min(\mathbf{a}, b, \delta_1, \dots, \delta_N), \\ g_j(\mathbf{a}^T \mathbf{x}_j + b) \geq 1 - \delta_j, \delta_j \geq 0, j = 1, \dots, N. \end{cases} \quad (1.1)$$

Обычно задачу (1) решают в двойственной форме:

$$\begin{cases} \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{j=1}^N \sum_{k=1}^N \left(g_j g_k \sum_{i=1}^n x_{ij} x_{ik} \right) \lambda_j \lambda_k \rightarrow \max(\lambda_1, \dots, \lambda_N), \\ \sum_{l=1}^n \lambda_j g_j = 0, \quad 0 \leq \lambda_j \leq C/2, \quad j = 1, \dots, N, \end{cases} \quad (1.2)$$

как задачу квадратичного программирования относительно неотрицательных множителей Лагранжа $\lambda_j, j = 1, \dots, N$. Связь коэффициентов направляющего вектора разделяющей гиперплоскости с множителями Лагранжа двойственной задачи:

$$a_i = \sum_{j=1}^N g_j \lambda_j x_{ij}, \quad i = 1, \dots, n. \quad (1.3)$$

Простота алгоритмической и численной реализации этого метода, явное указание на так называемые опорные элементы обучающей выборки (только они, по сути, и формируют разделяющую гиперплоскость), хорошие экспериментальные результаты сделали эту постановку задачи обучения распознавания образов наиболее популярной в последнее время. Именно поэтому такой критерий ляжет в основу построения метода отбора признаков в случае их упорядоченности в задаче обучения распознаванию образов.

Следует отметить, что постановка (1), является наиболее простым, так сказать, академичным из целого множества критериев, объединенных названием support vector machines. В таком варианте решение опирается на скалярные произведения

векторов признаков объектов распознавания и решающее правило является линейным в исходном пространстве. Существуют постановки с другим видом потенциальных функций (kernels), другими способами штрафов для случаев линейной неразделимости подвыборок двух классов. В литературе обсуждаются аспекты связи метода опорных векторов с методом потенциальных функций [1], методами беспризнакового распознавания [3]. Мы в наших предложениях для ясности рассуждений будем опираться на простую, ставшую уже канонической постановку (1).

2. РЕГУЛЯРИЗАЦИЯ В ОБУЧЕНИИ РАСПОЗНАВАНИЯ СИГНАЛОВ С УЧЕТОМ КРИТЕРИЯ ГЛАДКОСТИ РЕШАЮЩЕГО ПРАВИЛА

Для того, чтобы в процессе обучения предпочтение отдавалось решающим правилам с плавным изменением коэффициентов линейной части, в [2] предлагалось внести в критерий (1) дополнительную аддитивную составляющую $J'(\mathbf{a}) = \sum_{i=2}^n (a_i - a_{i-1})^2$. Нетрудно убедиться, что такая квадратичная функция может быть записана в виде $J'(\mathbf{a}) = \mathbf{a}^T \mathbf{B} \mathbf{a}$, где матрица \mathbf{B} ($n \times n$) имеет вид:

$$\mathbf{B}(n \times n) = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix}$$

Тогда целевая функция поиска параметров оптимальной разделяющей гиперплоскости, в отличие от (1), будет включать в себя еще одно слагаемое:

$$\mathbf{a}^T (\mathbf{I} + \gamma \mathbf{B}) \mathbf{a} + C \sum_{j=1}^N \delta_j \rightarrow \min(\mathbf{a}, b, \delta_1, \dots, \delta_N) \quad (2.1)$$

при тех же ограничениях. Здесь и далее коэффициент $\gamma \geq 0$ определяет степень, с которой штраф на негладкость последовательности коэффициентов участвует в процессе обучения. Очевидно, что матрица положительно определена.

Критерий в двойственной форме также практически не претерпит изменений, немного подкорректируется подсчет матрицы квадратичной формы:

$$\begin{cases} \sum_{j=1}^N \lambda_j - \frac{1}{2} \sum_{j=1}^N \sum_{k=1}^N (g_j g_k \mathbf{x}_j^T (\mathbf{I} + \gamma \mathbf{B})^{-1} \mathbf{x}_k) \lambda_j \lambda_k \rightarrow \max(\lambda_1, \dots, \lambda_N), \\ \sum_{j=1}^N \lambda_j g_j = 0, \quad 0 \leq \lambda_j \leq C/2, \quad j = 1, \dots, N. \end{cases} \quad (2.2)$$

Направляющий вектор оптимальной разделяющей гиперплоскости будет выражаться следующей формулой:

$$\mathbf{a} = (\mathbf{I} + \gamma \mathbf{B})^{-1} \sum_{j=1}^N \lambda_j g_j \mathbf{x}_j. \quad (2.3)$$

3. ОТБОР ПРИЗНАКОВ, ОСНОВАННЫЙ НА КОМБИНИРОВАНИИ ПОТЕНЦИАЛЬНЫХ ФУНКЦИЙ

В работах [5, 6] был предложен способ комбинирования потенциальных функций в задачах беспризнакового обучения распознаванию образов и восстановления числовой регрессии. Было показано, что этот механизм может быть перенесен на случай, когда объекты представлены своими векторами признаков, и достаточно эффективно применен как процедура неитерационного отбора информативных признаков.

В этом случае в «классический» критерий В.Н.Вапника (1) вводятся неотрицательные веса $r_1 \geq 0, \dots, r_n \geq 0$ при каждой компоненте направляющего вектора искомой разделяющей гиперплоскости:

$$\begin{cases} \sum_{i=1}^n [(1/r_i)a_i^2 + \log r_i] + C \sum_{j=1}^N \delta_j \rightarrow \min(\mathbf{a}, b, r_1, \dots, r_n, \delta_1, \dots, \delta_N), \\ g_j \left(\sum_{i=1}^n a_i x_i(\omega_j) + b \right) \geq 1 - \delta_j, \quad \delta_j \geq 0, \quad j = 1, \dots, N, \\ r_i \geq 0, \quad i = 1, \dots, n. \end{cases} \quad (3.1)$$

Двойственная задача квадратичного программирования:

$$\begin{cases} \sum_{i=1}^N \lambda_i - \frac{1}{2} \sum_{j=1}^N \sum_{k=1}^N \left(g_j g_k \sum_{i=1}^n r_i x_{ij} x_{ik} \right) \lambda_j \lambda_k \rightarrow \max(\lambda_1, \dots, \lambda_N), \\ \sum_{l=1}^N \lambda_l g_l = 0, \quad 0 \leq \lambda_j \leq C/2, \quad j = 1, \dots, N. \end{cases} \quad (3.2)$$

Для поиска параметров гиперплоскости и весовых коэффициентов используется метод Гаусса-Зайделя: фиксируется одна группа переменных $r_i \geq 0, i = 1, \dots, n$, проводится оптимизация по другой $\lambda_j, j = 1, \dots, N$ и т.д. Т.о. итерационный процесс обучения заключается в последовательном решении задачи квадратичного программирования (8) и пересчете весовых коэффициентов:

$$r_i^{next} = r_i^2 \sum_{j=1}^N \sum_{k=1}^N g_j g_k x_{ij} x_{ik} \lambda_j \lambda_k = a_i^2. \quad (3.3)$$

Направляющий вектор разделяющей гиперплоскости определяется в этом случае как:

$$a_i = r_i \sum_{j=1}^N g_j \lambda_j x_{ij}, \quad i = 1, \dots, n. \quad (3.4)$$

Такая процедура в ходе экспериментальных исследований продемонстрировала тенденцию к быстрой сходимости и отбору малого, даже порой слишком малого числа

признаков с существенно отличными от нулевых значений весами r_i , $i = 1, \dots, n$. В работе [6] был предложен несколько другой критерий, отличный от (7):

$$\begin{cases} \sum_{i=1}^n \frac{1}{r_i} \left(a_i^2 + \frac{1}{\mu} \right) + C \sum_{j=1}^N \delta_j \rightarrow \min(r_1, \dots, r_n, a_1, \dots, a_n, b, \delta_1, \dots, \delta_N), \\ \prod_{i=1}^n r_i = 1, \quad g_j \left(\sum_{i=1}^n a_i x_{ij} + b \right) \geq 1 - \delta_j, \quad \delta_j \geq 0, \quad j = 1, \dots, N. \end{cases} \quad (3.5)$$

В такой постановке присутствует параметр пользователя $\mu > 0$, позволяющий управлять степенью отбора признаков, названный степенью селективности. Двойственный критерий такой задачи в точности совпадает с (8), а вот пересчете весовых коэффициентов необходимо осуществлять по более сложной формуле:

$$r_i^{next} = \frac{a_i^2 + (1/\mu)}{\prod_{q=1}^n \sqrt[n]{(a_q^2 + (1/\mu))}}, \quad i = 1, \dots, n. \quad (3.6)$$

4. ОТБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ ПРИ ОБУЧЕНИИ РАСПОЗНАВАНИЮ ОБРАЗОВ С УПОРЯДОЧЕННЫМИ ПРИЗНАКАМИ

В этом разделе представлен критерий построения решающего правила, объединяющий отбор информативных признаков и наложение ограничений на разницу весовых коэффициентов. Возьмем за основу критерий (11), внесем дополнительный квадратичный штраф на разницу соседних весовых коэффициентов $r_i \geq 0$, $i = 1, \dots, n$, отнесенную к их произведению. Постановка задачи обучения с отбором признаков в случае их упорядоченности на оси:

$$\begin{cases} \sum_{i=1}^n \frac{1}{r_i} \left(a_i^2 + \frac{1}{\mu} \right) + \gamma \sum_{i=2}^n \frac{(r_i - r_{i-1})^2}{r_i r_{i-1}} + C \sum_{j=1}^N \delta_j \rightarrow \min(a_1, \dots, a_n, b, r_1, \dots, r_n, \delta_1, \dots, \delta_N), \\ \prod_{i=1}^n r_i = 1, \quad g_j \left(\sum_{i=1}^n a_i x_{ji} + b \right) \geq 1 - \delta_j, \quad \delta_j \geq 0, \quad j = 1, \dots, N. \end{cases} \quad (4.1)$$

Предлагается решать задачу минимизации критерия (13) методом Гаусса-Зайделя, разделив переменные на две группы: первая $\mathbf{a} \in R^n, b \in R, \delta_j \geq 0, j = 1, \dots, N$, и вторая – $r_i \geq 0, i = 1, \dots, n$, и осуществлять пошаговую минимизацию критерия по одной группе параметров, при фиксированной второй.

При фиксированных коэффициентах $r_i \geq 0, i = 1, \dots, n$, задача фактически сводится к классической постановке задачи обучения распознаванию образов методом опорных векторов. Двойственный критерий такой задачи в точности совпадает с (8). Для поиска весовых коэффициентов на каждом шаге метода покоординатного спуска необходимо решить систему нелинейных уравнений относительно параметров $r_i \geq 0, i = 1, \dots, n$ и фиктивного параметра η :

$$\begin{cases} -\frac{(a_1)^2+(1/\mu)}{r_1} + \gamma \left(\frac{r_1}{r_2} - \frac{r_2}{r_1} \right) + \eta = 0, & i = 1, \\ -\frac{(a_i)^2+(1/\mu)}{r_i} + \gamma \left(-\frac{r_{i-1}+r_{i+1}}{r_i} + \frac{r_i}{r_{i-1}} + \frac{r_i}{r_{i+1}} \right) + \eta = 0, & i = 2, \dots, n-1, \\ -\frac{(a_n)^2+(1/\mu)}{r_n} + \gamma \left(-\frac{r_{n-1}}{r_n} + \frac{r_n}{r_{n-1}} \right) + \eta = 0, & i = n, \\ \prod_{i=1}^n r_i - 1 = 0. \end{cases} \quad (4.2)$$

Каждое уравнение в этой системе включает всего 3-4 неизвестных, решение возможно осуществить методом простых итераций. Коэффициенты направляющего вектора разделяющей гиперплоскости вычисляются по правилу (10).

Следует отметить, что в таком подходе априорная информация об упорядоченности признаков налагает ограничения на весовые коэффициенты, отвечающие за информативность признаков, а не на сами коэффициенты направляющего вектора разделяющей гиперплоскости в отличие, например как в процедуре, описанной в [7].

ЗАКЛЮЧЕНИЕ

Статья демонстрирует способ совмещения в одном критерии отбора информативных признаков и наложения априорных, разумных с точки зрения решаемой задачи, ограничений на такой отбор. Сделан обобщающий обзор ранних публикаций авторов, из которых фактически и следует предложенная в статье идея. Выписан критерий и предложена схема его численной оптимизации. Необходимо всесторонне рассмотреть поведение предложенного алгоритма селективного отбора признаков в экспериментах как на модельных, так и на реальных задачах. Работа выполнена при поддержке грантов РФФИ № № 06-07-89249, 06-01-00412, 08-01-00695, 08-01-99003.

СПИСОК ЛИТЕРАТУРЫ

1. *Vapnik, V.* Statistical Learning Theory. New York: Wiley, 1998.
2. *O.S. Seredin, S.D. Dvoenko, O.V. Krasotkina, and V.V. Mottl* Machine Learning for Signal Recognition by the Criterion of Decision Rule Smoothness. Pattern Recognition and Image Analysis, Vol. 11, No. 1, 2001, pp. 87-90.
3. *Mottl V.V., Seredin O.S., Dvoenko S.D., Kulikowski C.A, Muchnik I.B.* Featureless pattern recognition in an imaginary Hilbert space. In: Proceedings of 16th International Conference Pattern Recognition, ICPR-2002, Quebec City, Canada, August, 2002, vol.II, pp.88-91.
4. *Mottl V.V., Seredin O.S., Krasotkina O.V., and Muchnik I.B.* Fusing of potential functions in reconstructing dependences from empirical data In: Doklady Mathematics, Vol. 71, No. 2, 2005, pp. 315-319. From Doklady Akademii Nauk, Vol. 401, No. 5, 2005, pp. 607-612.
5. *Mottl V.V., Seredin O.S., Krasotkina O.V., and Muchnik I.B.* Principles of multi-kernel data mining. In: P. Perner and A. Imiya (Eds.), Machine Learning and Data Mining in Pattern Recognition, Springer Verlag, LNAI 3587, 2005, pp. 52-61.
6. *Mottl V., Tatarchuk A., Sulimova V., Krasotkina O., and Seredin O.* Combining Pattern Recognition Modalities at the Sensor Level Via Kernel Fusion, In: Proceedings of 7th International Workshop Multiple Classifiers Systems, Prague, Czech Republic, 2007, pp. 1-12.
7. *O. Seredin, V. Mottl* Regularization in image recognition: the principle of decision rule smoothing, Pattern Recognition and Information Processing: Proceedings of the Ninth International Conference - Minsk, 2007. Vol.II., pp. 151-155.

Статья поступила в редакцию 04.05.2008