Визуализация классов в интеллектуальных системах на основе распознающих процедур

А. П. Немирко[1], Л. А. Манило[2]
Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)
[1]apn-bs@yandex.ru, [2]lmanilo@yandex.ru

***Аннотация*. Рассматриваются интеллектуальные системы поддержки принятия решений. Ставится задача найти такое отображение классов на плоскость, при котором удаленность классов друг от друга была бы максимальной. Для решения этой задачи применяются разные распознающие процедуры, в том числе и расстояние между выпуклыми оболочками классов. Введение добавочных весовых векторов в дискриминантный анализ Фишера позволяет отобразить классы на плоскости и более точно представить взаимное их положение в многомерном пространстве. Рассмотрена мера близости классов на основе оценки степени пересекаемости их выпуклых оболочек. Показано, что использование пересекаемости выпуклых оболочек может выявить потенциально достижимую точность классификации.**

Ключевые слова: системы поддержки принятия решений; линейный дискриминант Фишера; выпуклые оболочки; пересекаемость классов

I. Введение

Методы машинного обучения широко используются при разработке алгоритмов принятия решений для медицинской диагностики. При реализации многих интеллектуальных измерительных систем, таких как мониторные системы контроля состояния пациента важно найти небольшое число признаков и простые (линейные) решающие правила, реализуемые в системах реального времени. Визуализация классов на плоскости путем их отображения на две ортогональные оси может облегчить нахождение наилучшего линейного решающего правила. Для сокращения размерности признакового пространства часто применяется метод главных компонент (PCA) [1]. Для классификации используется также линейный дискриминант Фишера (ЛДФ) [2], который уменьшает размерность признакового пространства с исходного до одного путем проектирования многомерных данных на прямую. Экспериментальные исследования показывают, что критерий Фишера далеко не всегда оптимален для решения задачи распознавания объектов. Введение дополнительного весового вектора [3, 4] может уменьшить пересекаемость классов и привести к более эффективным процедурам линейной классификации на плоскости. Такой подход требует введения других, отличных от критерия Фишера, способов измерения близости классов в многомерном пространстве.

В данной работе предложен другой способ оценки близости классов и степени их пересекаемости. Необходимо отметить, что во всех случаях критерий Фишера остается универсальным.

II. Трансформация признакового пространства на основе критерия Фишера

Для задачи классификации на два класса в многомерном признаковом пространстве линейная разделяющая граница определяется выражением $\mathbf{W^T X} - a = 0$, где $\mathbf{W}$ – вектор весовых коэффициентов, $\mathbf{X}$ – входной вектор, a – скалярная пороговая величина.

Линейный дискриминант Фишера определяется как вектор $\mathbf{W}$, для которого линейный функционал

$$J(\mathbf{W}) = (m_1 - m_2)^2 \big/ (s_1^2 + s_2^2) \qquad (1)$$

максимален. В этой формуле m_1 и m_2 - средние значения классов, спроектированных на $\mathbf{W}$, s_1^2 и s_2^2 - выборочные внутриклассовые рассеяния для этих проекций. Для $\mathbf{W}$ при условии, что $J(\mathbf{W})$ = max, расстояние между проекциями классов на $\mathbf{W}$ максимально.

Приведенный выше, функциональный критерий (1) можно переписать в виде

$$J(\mathbf{W}) = \frac{\mathbf{W}^T \mathbf{S}_B \mathbf{W}}{\mathbf{W}^T \mathbf{S}_W \mathbf{W}}, \qquad (2)$$

где $\mathbf{S}_B = (\mathbf{M}_1 - \mathbf{M}_2)(\mathbf{M}_1 - \mathbf{M}_2)^T$, — матрица межклассового рассеяния, $\mathbf{M}_1$ и $\mathbf{M}_2$ — векторы средних значений классов; $\mathbf{S}_W = \mathbf{S}1 + \mathbf{S}2$ — матрица внутриклассового рассеяния; $\mathbf{S}1$, $\mathbf{S}2$ — матрицы внутриклассового рассеяния 1-го и 2-го классов,

Работа выполнена при финансовой поддержке РФФИ, проекты № 18-07-00264 и № 16-01-00159

$$\mathbf{S}1 = \sum_{i=1}^{n_1} (\mathbf{X}_i^{(1)} - \mathbf{M}_1)(\mathbf{X}_i^{(1)} - \mathbf{M}_1)^T,$$
$$\mathbf{S}2 = \sum_{i=1}^{n_2} (\mathbf{X}_i^{(2)} - \mathbf{M}_2)(\mathbf{X}_i^{(2)} - \mathbf{M}_2)^T,$$

$\mathbf{X}_i^{(j)}$ – i-й входной вектор j-го класса, n_1 и n_2 – число членов каждого класса.

Анализ формулы (2) показывает [2], что максимум $J(\mathbf{W})$ достигается при

$$\mathbf{W} = \mathbf{S}_W{}^{-1}(\mathbf{M}_1 - \mathbf{M}_2). \quad (3)$$

Для двух классовой задачи распознавания дискриминантный анализ Фишера дает лишь оптимальный весовой вектор **W** при максимуме критерия Фишера. Если его дополнить процедурой нахождения порогового значения a, то он может использоваться и для классификации на два класса. Однако не всегда в этом случае мы получим наилучшие результаты классификации. В работах [3,4] показано, что применение добавочных весовых векторов в этой задаче может улучшить качество классификации, в частности, иногда сделать классы линейно разделимыми.

Для исходного n-мерного признакового пространства мы можем переписать выражение (3) в виде

$$\mathbf{W}_n = \mathbf{S}_n^{-1}(\mathbf{m1}_n - \mathbf{m2}_n) \quad (4)$$

В работе [3] выведена формула для рекуррентного вычисления других дополнительных ортогональных весовых векторов. Второй дополнительный весовой вектор $\mathbf{W}_{n-1}$ в (n-1)-мерном пространстве определяется выражением [3]

$$\begin{aligned} \mathbf{W}_{n-1} &= [\mathbf{S}_n + \mathbf{W}_n{}^T \mathbf{S}_n \mathbf{W}_n (\mathbf{W}_n \mathbf{W}_n{}^T)]^{-1} \mathbf{B} \\ \mathbf{B} &= [(\mathbf{m1}_n - \mathbf{m2}_n) - \mathbf{W}_n{}^T (\mathbf{m1}_n - \mathbf{m2}_n) \mathbf{W}_n] \end{aligned}. \quad (5)$$

Далее будем обозначать весовой вектор, найденный, по формуле (4), через **W**1, а по формуле (5) через **W**2.

На экспериментальных результатах [3] было показано, что использование только одного весового вектора, в примере с непересекающимися классами не обеспечивает линейной разделимости классов. Добавочный же признак обеспечивает полную линейную разделимость этих классов.

III. Использование выпуклых оболочек

Можно оценить эффективность введения добавочных весовых векторов, если применить другие, отличные от критерия Фишера, способы измерения близости классов в многомерном пространстве. Критерий Фишера «измеряет» насколько далеко классы расположены друг от друга. Для другого способа измерения близости классов определим выпуклую оболочку множества точек в n-мерном евклидовом пространстве, как наименьшее выпуклое множество, содержащее все эти точки, и для двухклассовой задачи найдем область пересечения двух выпуклых оболочек рассматриваемых классов. Первый параметр, характеризующий близость двух классов друг к другу, оценивает степень пересекаемости их выпуклых оболочек. В случае полной непересекаемости этот параметр g (%) равен нулю. При наличии пресечения классов g равно доли числа точек обоих классов обучающей выборки, попавших в зону пересечения, т.е.

$$g = (n_1 + n_2)/(N_1 + N_2), \quad (6)$$

где n_1, n_2 – число точек 1-го и 2-го классов, попавших в зону пересечения выпуклых оболочек; N_1, N_2 – число членов обучающей выборки 1-го и 2-го классов. Очевидно, что $0 < g < 100\,\%$.

IV. Использование методов вычислительной геометрии

При решении задачи оценки пересекаемости классов используются методы вычислительной геометрии [5, 6]. В нашем случае задача определения пересечения выпуклых оболочек близка задаче обнаружения столкновений геометрических тел (collision detections) в компьютерной геометрии, которая часто используется в машинном зрении и компьютерных играх. В системе MATLAB для этой цели используются функции построения выпуклой оболочки множества точек в многомерном пространстве: `delaunayn`, `convhulln` [6]; функция обнаружения проникновений множества точек в заданную выпуклую оболочку в 2D пространстве `inpolygon` [8], и в многомерном пространстве: `tsearchn` [7], `inhull` [9]. Эти средства обеспечивают решение следующих задач:

- Обнаружение линейно непересекающихся классов в многомерном пространстве.
- Оценку степени пересекаемости классов в многомерном пространстве путем вычисления параметра g.
- Сравнение эффективности линейных преобразований в исходном признаковом пространстве с позиций пересекаемости классов (по величине g).
- Построение более эффективных линейных решающих правил после трансформаций исходного пространства на основе алгоритмов вычислительной геометрии.

Однако решения этих задач недостаточно для построения оптимальных решающих правил, но данные, полученные с их помощью, облегчают представление о расположении классов в многомерном пространстве и обеспечивают их визуализацию на плоскости.

V. Проведенные эксперименты

Для иллюстрации сказанного возьмем два класса ирисов Фишера: виргинского (virginica) и разноцветного

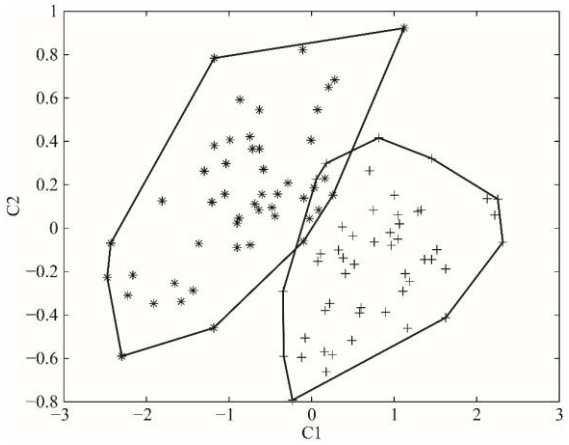

Рис. 1. Применение метода PCA для двух классов ирисов Фишера: разноцветного и виргинского. На рисунке также показаны выпуклые оболочки (полигоны) этих классов

(versicolor) из пакета MATLAB, которые измерены по 4 признакам: длина и ширина чашелистика; длина и ширина лепестка. Каждый класс содержит 50 членов [10]. Применение в системе MATLAB к этим классам метода главных компонент (PCA) [1] дает распределения точек на плоскости (рис. 1). Использование функции `convexHull` дает возможность на этом же рисунке показать выпуклые оболочки данных классов (полигоны), а функция `inpolygon` позволяет увидеть и подсчитать взаимные проникновения точек классов в выпуклые оболочки противоположных классов. Расчеты дают следующие результаты: $n_1 = 2$, $n_2 = 6$, и по формуле (6) $g = (2 + 6)/100 = 8\%$.

Далее определим оптимальный весовой вектор **W**1 согласно линейному дискриминанту Фишера и вычислим добавочный весовой вектор **W**2 по формуле (5). Тогда на плоскости этих двух векторов классы и соответствующие выпуклые оболочки будут иметь вид рис. 2. Согласно рис. 2, рис. 3 и расчетам по функции `inpolygon` в этом случае $n_1 = 1$, $n_2 = 2$, $g = (1 + 2)/100 = 3\%$.

После применения к исходным описаниям классов ирисов Фищера в 4-мерном пространстве функций `convhulln` пересекаемость классов в этом 4-мерном пространстве определяется следующими значениями параметров: $n_1 = 0$, $n_2 = 1$, $g = (1+0)/100=1\%$. В результате применения методов компьютерной геометрии мы выяснили, что только один представитель 1-го класса 'versicolor' попал в выпуклую оболочку 2-го класса 'virginica'. Программа вычислила, что это 34-й элемент массива. Число же элементов 'virginica', попавших в выпуклую оболочку 'versicolor', оказалась равной нулю.

Полученные результаты говорят о следующем.

1. Рассмотренные множества линейно неразделимы в признаковом пространстве. Потенциально достижимая пересекаемость классов на данной обучающей выборке составляет $g = 1\%$.

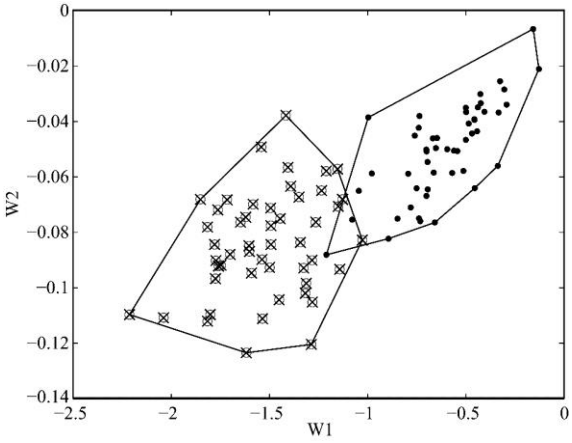

Рис. 2. Расположение классов ирисов Фишера в преобразованном пространстве: **W**1 – оптимальный весовой вектор, найденный при максимизации критерия Фишера (1); **W**2 – добавочный весовой вектор, найденный по формуле (5) согласно [2]. На рисунке справа изображен 1-й класс ‘versicolor’

2. Применение метода PCA обладает небольшой эффективностью с точки зрения пересекаемости классов ($g = 8\%$).

3. Применение метода ЛДФ с одним добавочным признаком уменьшает пересекаемость классов до $g = 3\%$.

Заменив 34-й элемент 1-го класса немного измененной копией 1-го элемента того же класса мы получаем два непересекающихся класса. Применив к ним персептронное обучение, получим весовой вектор $\mathbf{P}_n = (0.34, 0.36, -0.55, -0.52, 0.43)$, полностью разделяющий эти классы. Его последняя компонента связана с пороговой величиной. Взяв только первые 4 компоненты этого вектора получим $\mathbf{W}_p = (0.34, 0.36, -0.55, -0.52)$.

Далее найдем проекции наших классов на данный весовой вектор и проверим эти проекции, сравнив их с порогом 0.43. Тогда решающее правило, полностью разделяющее эти два класса имеет вид:

Если $\mathbf{W}_p^T\mathbf{X} \le 0.43$, то мы имеем 1-й класс, иначе решаем, что это 2-й класс. Здесь $\mathbf{W}_p^T\mathbf{X} = 0.34\,x_1 + 0.36\,x_2 - 0.55\,x_3 - 0.52\,x_4$, $\mathbf{W}_p$ — весовой вектор-столбец, найденный в результате персептронного обучения, $\mathbf{X} = (x_1, x_2, x_3, x_4)$ — входной вектор-столбец.

Попытаемся как-то визуализировать результат пересечения классов в 4-мерном пространстве. Для этого построим плоскость из двух весовых векторов. Первый это полученный вектор $\mathbf{W}_p = \mathbf{V}1$, а второй $\mathbf{V}2$ строится как перпендикуляр к $\mathbf{V}1$ и находится как наилучший весовой вектор в перпендикулярной плоскости, которая является (n-1)-мерном пространством, по максимуму критерия Фишера. Тогда мы получим расположение классов, показанное на рис. 4. Этот рисунок наглядно показывает,

что только один элемент из двух множеств классов входит в зону пересечения двух рассматриваемых классов.

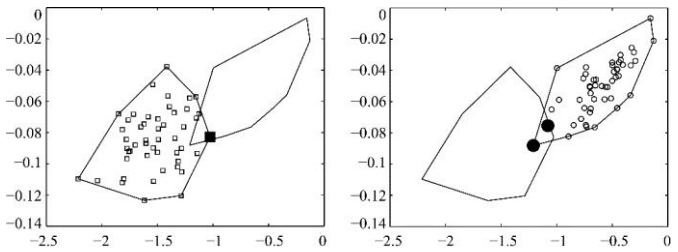

Рис. 3. Взаимное проникновение точек классов в выпуклые оболочки противоположных классов в пространстве двух найденных весовых векторов **W1** и **W2**, рассчитанное по функции `inpolygon`. На рисунке классы и оси те же, что и на рис. 2.

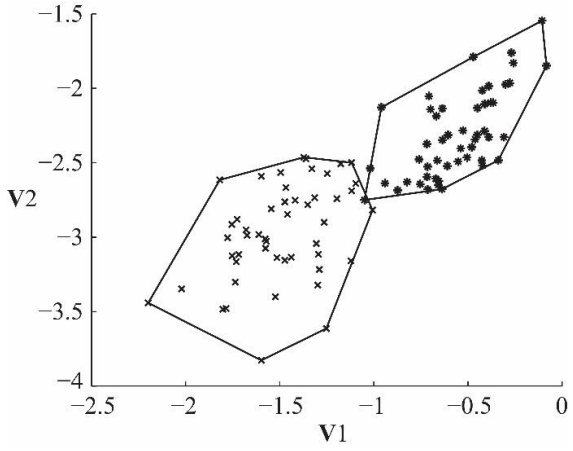

Рис. 4. Визуализация расположения классов ирисов Фишера: Расположение классов ирисов Фишера в преобразованном пространстве: **V**1 – оптимальный весовой вектор, найденный при персептронном обучении непересекающихся классов; **V**2 – добавочный весовой вектор. На рисунке справа изображен 1-й класс – ‘versicolor’

VI. Выводы и предложения

Проведенные экспериментальные исследования продемонстрировали полезность применения методов компьютерной геометрии при поиске наилучшего решающего правила в 4-мерном пространстве признаков. В частности удалось вычислить потенциально достижимую минимальную пересекаемость обучающих выборок при использовании линейных преобразований признакового пространства. К сожалению, функции построения выпуклых оболочек, триангуляции и поиска взаимного проникновения точек классов в выпуклые оболочки других классов хорошо работают только при небольшом числе признаков.

Существует и другой подход к оценке близости классов друг к другу в виде минимального расстояния между двумя выпуклыми оболочками этих классов. Сложность измерения такого расстояния возрастает с увеличением размерности пространства. Существует несколько алгоритмов измерения такого минимального расстояния для 2D и 3D мерных случаев. Можно использовать упрощенный способ, который заключается в перемещении одной из оболочек в направлении вектора, соединяющего центры классов до момента, когда все элементы будут удалены из области пересечения (или до момента, когда первый элемент окажется в области пересечения для непересекающихся классов). Расстояние полученного сдвига и будет искомым расстоянием.

Список литературы

[1] Jolliffe, I.T.: Principal Component Analysis. 2nd ed., New York: Springer-Verlag, 2002. 487 p.

[2] Duda R.O., Hart P.E., Stork D.G.: Pattern Classification (Pt.1). New York: Wiley, 2001. 659 p.

[3] Nemirko A.P.: Transformation of feature space based on Fisher’s linear discriminant. Pattern Recognition and Image Analysis, vol. 26(2), pp.:257–261 (2016). doi: 10.1134/S1054661816020127

[4] Manilo L.A., Nemirko A.P.: Recognition of biomedical signals based on their spectral description data analysis. Pattern Recognition and Image Analysis, vol. 26(4), pp. 782–788 (2016). doi: 10.1134/S1054661816040088

[5] Preparata F.P., Shamos M.I.: Computational Geometry: An Introduction. Springer-Verlag, (1985)

[6] Berg M., Cheong O., Kreveld M., Overmars M.: Computational Geometry: Algorithms and Applications. Third Edition. Springer-Verlag, (2008).

[7] Barber, C.B., Dobkin, D.P., and Huhdanpaa, H.T.: The Quickhull algorithm for convex hulls. ACM Trans. on Mathematical Software, 22(4):469–483, (1996). http://www.qhull.org

[8] Inpolygon. MathWorks Documentation (R2016b). MATLAB Function Reference (2016). http://www.mathworks.com/help/matlab/functionlist.html?requestedDomain=www.mathworks.com

[9] Inhull by John D'Errico. Efficient test for points inside a convex hull in n dimensions. MathWorks, File Exchange. (2009). http://www.mathworks.com/matlabcentral/fileexchange/10226-inhull

[10] Iris Data Set. UCI Machine Learning Repository (2016). https://archive.ics.uci.edu/ml/datasets/Iris