УДК 51.76, 57.087

Линейный дискриминантный анализ Фишера в задачах классификации многомерных биомедицинских данных

А.П. Немирко

Анатолий Павлович Немирко –
д.т.н., профессор,
кафедра биотехнических систем,
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

E-mail: apn-bs@yandex.ru

Исследовано линейное преобразование пространства на основе критерия Фишера для задач классификации многомерных биомедицинских данных. Приведены выражения для рекуррентного вычисления весовых векторов, которые образуют новые признаки. Показано, что за счет более точного представления данных с помощью найденных признаков они позволяют более эффективно реализовать процедуры классификации.

Ключевые слова: *анализ биомедицинских данных, линейный дискриминант Фишера, близость классов.*

Linear transformation of space on the basis of Fischer's criterion for problems of multidimensional biomedical data classification is investigated. Expressions for recurrent calculation of weight vectors which form new features are given. It is shown that due to more exact data presentation by means of the found features they allow to realize classification procedures more effectively.

Keywords: *biomedical data analysis, liner Fisher's discriminant, classes proximity.*

Классификация многомерных данных широко применяется в биологических и медицинских исследованиях [1, 2]. Для сокращения размерности признакового пространства часто применяется метод главных компонент [3]. Для классификации используется также линейный дискриминант Фишера (ЛДФ) [4], 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 [5–7] 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 предложено несколько способов оценки близости классов и степени их пересекаемости. Все эти способы применены к описанию классов в пространстве двух весовых векторов, найденных по критерию Фишера. Необходимо отметить, что во всех случаях критерий Фишера остается универсальным.

Цель работы – исследование применимости дискриминантного анализа с дополнительными весовыми векторами для повышения качества распознавания многомерных биомедицинских данных.

Трансформация признакового пространства на основе критерия Фишера

Линейный дискриминант Фишера определяется как вектор **W,** для которого линейный функционал

$$J(\mathbf{W}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}, \quad (1)$$

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

Приведенный выше функциональный критерий (1) можно переписать в виде

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

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- и 2-го классов:

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

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

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

Анализ формулы (2) показывает [4], что максимум $J(\mathbf{W})$ достигается при

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

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

$$\mathbf{W}_1 = \mathbf{S}_1^{-1}(\mathbf{m}1_1 - \mathbf{m}2_1). \quad (4)$$

В работе [5] выведена формула для рекуррентного вычисления других дополнительных ортогональных весовых векторов. Второй дополнительный весовой вектор $\mathbf{W}_2$ определяется выражением

$$\mathbf{W}_2 = [\mathbf{S}_1 + \mathbf{W}_1{}^T \mathbf{S}_1 \mathbf{W}_1 (\mathbf{W}_1 \mathbf{W}_1{}^T)]^{-1} \times$$
$$\times [(\mathbf{m}1_1 - \mathbf{m}2_1) - \mathbf{W}_1{}^T (\mathbf{m}1_1 - \mathbf{m}2_1)\mathbf{W}_1]. \quad (5)$$

Эксперименты с линейно разделимыми классами

Для экспериментальных исследований выбраны два линейно разделимых многомерных множеств f1 и f2 [5]. Результат вычисления по ним первых двух главных компонент показан на рис. 1. Из него видно, что метод главных компонент не обеспечивает линейную разделимость классов.

Реализация метода ЛДФ с одним добавочным признаком на данных множествах показана на рис. 2.

Из рис. 2 видно, что использование только одного весового вектора $\mathbf{W}_1$, найденного по критерию Фишера, линейной разделимости достичь не удается. Добавочный же признак (весовой вектор $\mathbf{W}_2$) обеспечивает полную линейную разделимость классов f1 и f2.

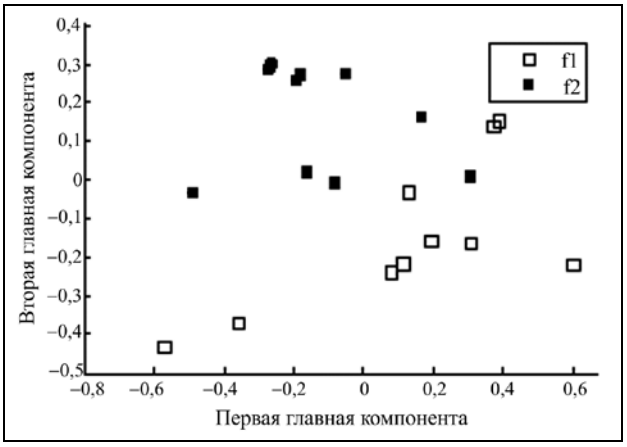

Рис. 1. Анализ множеств f1 и f2 по методу главных компонент

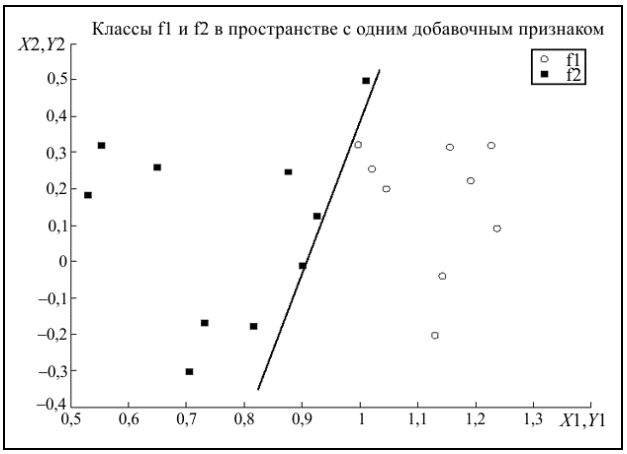

Рис. 2. Анализ множеств f1 и f2 методом ЛДФ с одним добавочным признаком. Ось абсцисс – вектор $\mathbf{W}_1$, ось ординат – вектор $\mathbf{W}_2$

На рис. 3 изображена область пересечений двух классов ирисов Фишера: виргинского и разноцветного справа. [9] в сокращенном пространстве, образованном двумя весовыми векторами $\mathbf{W}_1$ и $\mathbf{W}_2$.

Из рис. 3 видно, что только в двумерном пространстве можно достичь нулевой ошибки классификации виргинских ирисов при малых ошибках классификации разноцветных ирисов.

В случае сложных конфигураций распределений классов выигрыш от введения дополнительного весового вектора может оказаться существенным. На рис.4 показаны два класса линейно непересекающихся объектов, представленных в двумерном пространстве признаков $x1$ и $x2$. Для данных распределений этих классов использование критерия Фишера дает решающий вектор $\mathbf{W}_1$, показанный на рисунке, с общей ошибкой классификации, равной 33%. На плоскости же легко найти решающий вектор **F**, который разделяет эти классы безошибочно.

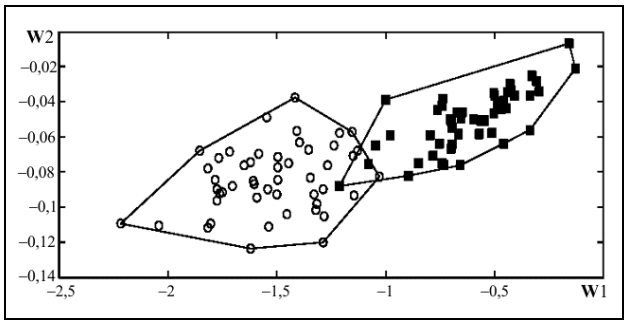

Рис. 3. Анализ ирисов Фишера: виргинского (слева) и разноцветного (справа) в редуцированном пространстве методом ЛДФ с одним добавочным признаком

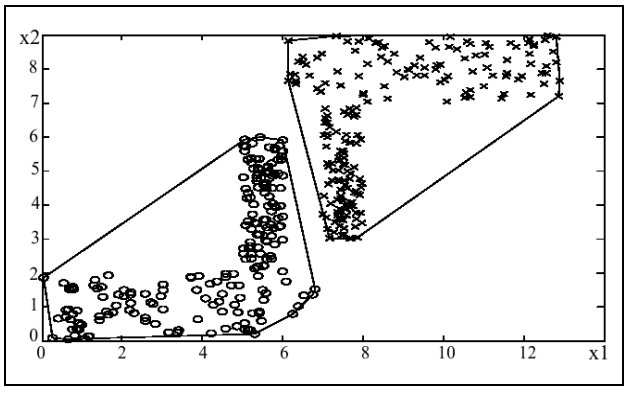

Рис. 4. Классификация при сложной конфигурации распределений: **W**1 – оптимальный решающий вектор, найденный по критерию Фишера; **F** – решающий вектор для безошибочной классификации на 2 класса

Оценка близости классов в многомерном пространстве

Экспериментальные исследования показывают, что критерий Фишера далеко не всегда оптимален для решения задачи распознавания объектов. Для улучшения классификации нужны новые критерии близости классов. Эти критерии не всегда легко применить в исходном пространстве. Однако можно сначала снизить размерность признакового пространства с помощью ЛДФ с добавочными признаками, а затем искать оптимальный весовой вектор уже в сокращенном пространстве, например, в двумерном. Этот подход требует введения других, отличных от критерия Фишера, способов измерения близости классов в двумерном пространстве.

Можно предложить несколько способов оценки близости классов и степени их пересекаемости. Необходимо отметить, что во всех случаях критерий Фишера остается универсальным.

1. П е р в ы й с п о с о б оценки пересекаемости классов для 2-классовой задачи заключается в измерении размера области пересечения, которую можно получить построив выпуклые оболочки обоих классов и найдя область их пересечения. Степень близости классов (при пересечении степень пересекаемости) может быть оценена как минимальное расстояние между этими выпуклыми оболочками. Существует несколько алгоритмов измерения такого минимального расстояния (например, алгоритм Гилберта – Джонсона – Кёрти [8]). Мы применяем способ, который заключается в передвижении одной из оболочек в направлении вектора, соединяющего центры классов до момента, когда все элементы будут удалены из области пересечения (или до момента, когда 1-й элемент окажется в области пересечений для непересекающихся классов). Расстояние полученного сдвига и будет искомым расстоянием. Алгоритм реализуется в системе MATLAB с применением функций convhull, convhulln.

2. В т о р о й с п о с о б оценки пересекаемости классов заключается в подсчете доли элементов каждого класса, попавших в область пересечения, и вычислении среднего по двум классам. Этот параметр меняется от 0 до 1. Конечно, для задачи классификации важно, сколько представителей каждого класса попало в область пересечений. Это влияет на ошибки классификации.

3. С точки зрения классификации наиболее адекватным следует считать третий способ, который измеряет площадь под кривой зависимости доли правильно обнаруженных элементов одного класса от доли ошибочного обнаружения второго класса – аналог кривой рабочей характеристики (ROC-кривой). Методы преобразования пространства признаков можно сравнивать по этому критерию. Чем площадь больше – тем метод лучше.

- Для классификации многомерных биомедицинских данных используется линейный дискриминант Фишера. Экспериментальные исследования показали, что критерий Фишера не всегда оптимален для решения задачи распознавания объектов. Введение дополнительных весовых векторов уменьшает пересекаемость классов и приводит к более эффективным процедурам линейной классификации.

Найдено рекуррентное выражение для последовательного вычисления добавочных признаков.

Предложено несколько способов оценки близости классов и степени их пересекаемости, которые могут быть использованы для реализации процедур классификации.

Исследование выполнено при поддержке РФФИ в рамках научных проектов №№ 15-07-01790, 16-01-00159 и медицинского проекта CardioQVARK – кардиограмма с помощью телефона www.cardio-qvark.ru).

🕮 Литература

1. *Немирко А.П., Манило Л.А., Калиниченко А.Н.* Математические методы анализа биомедицинских данных. СПб: Изд-во СПбГЭТУ «ЛЭТИ». 2013. 175 с.
2. *Рангайян Р.М.* Анализ биомедицинских сигналов. Практический подход / Пер. с англ. *А.Н. Калиниченко* / *Под ред. А.П. Немирко*. М.: Физматлит. 2007. 440 с.
3. *Айвазян С.А., Бежаева З.И., Староверов О.В.* Классификация многомерных наблюдений. М.: Статистика. 1974. 240 с.
4. *Дуда Р., Харт П.* Распознавание образов и анализ сцен / Пер. с англ. М.: Мир. 1976. 511 с.
5. *Nemirko A.P.* Transformation of feature space based on Fisher's linear discriminant. // Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications. 2016. V. 26. № 2. P. 257–261.
6. *Манило Л.А.* Упорядочение спектральных признаков по эмпирическим оценкам межгруппового расстояния в задачах классификации биосигналов // Известия вузов России. Радиоэлектроника. 2006. Вып. 3. С. 20–29.
7. *Манило Л.А.* Линейный дискриминант Фишера в задачах распознавания биосигналов по частотным свойствам // Сб. докл. 12-й Всерос. конф. «Математические методы распознавания образов» (Москва, 2005). М.: МАКС Пресс. 2005. С. 371–374.
8. *Gilbert E.G., Johnson D.W., Keerthi S.S.* A fast procedure for computing the distance between complex objects in three-dimensional space // IEEE Journal of Robotics and Automation. April 1988. V. 4. P. 193–203.
9. Iris Data Set. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Iris, 2016.

Поступила 1 августа 2016 г.

Fischer's linear discriminant analysis in problems of multidimensional biomedical data classification

A.P. Nemirko
Dr.Sc. (Eng.), Professor, Department of Biotechnical Systems, Saint Petersburg Electrotechnical University «LETI»
E-mail: apn-bs@yandex.ru

For reduction of feature space dimension at biomedical data classification the Fischer's linear discriminant is used. It reduces space dimension from initial to one by projecting the multidimensional data into a straight line. Pilot studies show that Fischer's criterion is not always optimum for the solution of an object recognition problem. Introduction of an additional weight vector can reduce an intersection between classes and lead to more effective procedures of linear classification on the plane. Recurrent expression for consecutive calculation of additional features is derived. When only one additional weight vector is used, procedure

of classification is realized on the plane. Such approach demands introduction of other ways of class proximity measurements different from Fischer's criterion. In this work some ways of an assessment of class adjacency and degree of their intersection are offered. All these ways are applied to classification in space of two weight vectors found by Fischer's criterion.

REFERENCES

1. Nemirko A.P., Manilo L.A., Kalinichenko A.N. Matematicheskie metody analiza biomedicinskih dannyh. SPb: Izd-vo SPbGJeTU «LJeTI». 2013. 175 s.
2. Rangajjan R.M. Analiz biomedicinskih signalov. Prakticheskij podhod / Per. s angl. A.N. Kalinichenko / Pod red. A.P. Nemirko. M.: Fizmatlit. 2007. 440 s.
3. Ajvazjan S.A., Bezhaeva Z.I., Staroverov O.V. Klassifikacija mnogomernyh nabljudenij. M.: Statistika. 1974. 240 s.
4. Duda R., Hart P. Raspoznavanie obrazov i analiz scen / Per. s angl. M.: Mir. 1976. 511 s.
5. Nemirko A.P. Transformation of feature space based on Fisher's linear discriminant. // Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications. 2016. V. 26. № 2. P. 257–261.
6. Manilo L.A. Uporjadochenie spektral'nyh priznakov po jempiricheskim ocenkam mezhgruppovogo rasstojanija v zadachah klassifikacii biosignalov // Izvestija vuzov Rossii. Radiojelektronika. 2006. Vyp. 3. S. 20–29.
7. Manilo L.A. Linejnyj diskriminant Fishera v zadachah raspoznavanija biosignalov po chastotnym svojstvam // Sb. dokl. 12-j Vseros. konf. «Matematicheskie metody raspoznavanija obrazov» (Moskva, 2005). M.: MAKS Press. 2005. S. 371–374.
8. Gilbert E.G., Johnson D.W., Keerthi S.S. A fast procedure for computing the distance between complex objects in three-dimensional space // IEEE Journal of Robotics and Automation. April 1988. V. 4. P. 193–203.
9. Iris Data Set. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Iris, 2016.