Визуализация многомерных данных с использованием метода опорных векторов

А. П. Немирко
Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)
apn-bs@yandex.ru

***Аннотация*. Рассмотрена задача визуализации при линейном анализе двух классов в многомерном признаковом пространстве. В результате линейного преобразования координат находятся две ортогональные оси, при проектировании классов на которые классы максимально удалены друг от друга. Близость классов оценивается на основе критерия минимального расстояния между их выпуклыми оболочками. Такой критерий позволяет отображать случаи полной разделимости и случайных выбросов. Для получения ортогональных векторов редуцированного пространства применяется метод опорных векторов, который обеспечивает получение весового вектора, определяющего минимальное расстояние между выпуклыми оболочками классов для линейно разделимых классов. Для пересекающихся классов использованы алгоритмы с уменьшением, сжатием и сдвигом выпуклых оболочек. Экспериментальные исследования посвящены применению исследованных методов визуализации для анализа биомедицинских данных.**

Ключевые слова: визуализация многомерных данных; машинное обучение; метод опорных векторов; анализ биомедицинских данных

I. Введение

Несмотря на бурное развитие нейросетевого подхода к распознаванию образов остается обширная область приложений, для которых характерно описание классов в многомерном признаковом пространстве и поиск решений в этом пространстве. Особенно много таких задач в биологии и медицине. При их решении для исследователя важно знать насколько пересекаются классы, и попытаться построить наилучшую разделяющую поверхность. При работе в многомерном пространстве область пересечения классов невидима, а решающие правила строятся исходя из каких либо теоретических гипотез. Однако часто бывают случаи, когда более подробное изучение области пересечений имеет особую важность, например, в случае высоких значений стоимостей ошибок диагностики или при обнаружении выскакивающих точек, не вписывающихся в описание некоторого биологического вида. Возникает проблема адекватного отображения области пересечений в 2-мерном пространстве. Иначе эту проблему можно назвать визуализацией классов на плоскости.

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

Для этой цели обычно применяются статистические методы сокращения размерности и визуализации: метод главных компонент (PCA) [1] и метод отображения на плоскость [2, 3, 4], основанный на линейном дискриминанте Фишера (FDA) [5]. К сожалению, эти методы не дают исчерпывающей картины области пересечений классов и не обеспечивают отображение случаев полной разделимости или случайных выбросов [2].

При проектировании многомерных классов на плоскость из-за информационных потерь возникают ошибки, которые проявляются в появлении дополнительных экспериментальных точек в области пересечения классов. Здесь возникает задача, найти такое отображение классов на плоскость, чтобы число экспериментальных точек, попавших в область пересечения на плоскости, было бы такое же, как и в многомерном пространстве (меньшее их число невозможно).

В данной работе пересечение классов рассматривается как пересечение их выпуклых оболочек. Поэтому область пересечения классов рассматривается как область пересечения их выпуклых оболочек, как в многомерном пространстве, так и на плоскости. В качестве критерия преобразования пространства рассматривается минимальное расстояние между их выпуклыми оболочками. Тогда задача визуализации формулируется следующим образом. Найти такое подпространство размерности 2, в ортогональных проекциях на которые расстояние между выпуклыми оболочками классов было бы минимальным.

II. Использование распознающих процедур

Пусть $\mathbf{x}_i, i=1,2,...,N$ – это векторы в n-мерном признаковом пространстве обучающего множества X. Они принадлежат одному из двух классов ω_1, ω_2. Задача линейного распознавания заключается в том, чтобы найти гиперплоскость $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0$, которая бы оптимально классифицировала все вектора обучающего множества, где $\mathbf{w}$ – весовой вектор, w_0 – скалярная пороговая величина. Здесь $\mathbf{w}$ ищется так, чтобы гиперплоскость $g(\mathbf{x})$, которая перпендикулярна $\mathbf{w}$, наилучшим образом (в смысле критерия обучения) разделяла бы классы. Поэтому его можно рассматривать

как первый признак для визуализации классов на плоскости. Второй признак ищется на плоскости перпендикулярной **w** по критерию обучения (тому же или другому). Далее, в основном, рассматривается распознающая процедура метода опорных векторов [6. 14]. Известно, что в случае непересекающихся классов метод опорных векторов (SVM) вычисляет минимальное расстояние между выпуклыми оболочками классов и соответствующий ему весовой вектор **w** [15].

Для метода SVM при линейно разделимых классах задача нахождения оптимальной разделяющей гиперплоскости формулируется как

$$\begin{cases} \|\mathbf{w}\|^2 \to \min \\ y_i\left(\mathbf{w}^T\mathbf{x}_i + w_0\right) \ge 1, \quad i = 1,2,...,N \end{cases}, \tag{1}$$

где y_i – показатель класса для каждого $\mathbf{x}_i$, равный +1 для ω_1 и –1 для ω_2. Это задача выпуклого квадратичного программирования (по $\mathbf{w}, w_0$) в выпуклом множестве с учетом совокупности линейных неравенств. Ее решение имеет вид

$$\begin{aligned} &\mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \\ &w_0 = \mathbf{w}^T\mathbf{x}_i - y_i, \quad \lambda_i > 0 \end{aligned},$$

где λ_i - множители Лагранжа.

В общем виде проблему нахождения минимального расстояния между выпуклыми оболочками классов (NPP - nearest point problem) решают также другие алгоритмы: SK-алгоритм (Schlesinger–Kozinec algorithm) [7], MDM-алгоритм (Mitchell–Dem'yanov–Malozemov algorithm) [8]. Эти алгоритмы также могут быть использованы для получения весового вектора для визуализации классов.

III. Случай линейно неразделимых классов

В случае линейно неразделимых классов **A** и **B** вместо критерия минимального расстояния между их выпуклыми оболочками Conv(**A**) и Conv(**B**) можно использовать критерий минимальной глубины проникновения или глубины взаимного пересечения классов *D*, который используется в задачах обнаружения столкновений [9, 16]. Этот критерий определяется как минимальная величина, на которую надо сдвинуть **B** в любом направлении, чтобы Conv(**A**) не пересекалась бы с **B**. Нахождение такого направления дает требуемый вектор **w**. Следуя этой же методике можно заключить, что в случае разделимых классов расстояние между ними можно определить как минимальное расстояние, на которое надо сдвинуть Conv(**A**) в любом направлении, чтобы **A** и **B** пересекались (рис. 1).

Существуют алгоритмы нахождения глубины проникновения для 2D и 3D случаев [9, 16], однако для *n*D случая нахождение глубины проникновения представляет собой трудную задачу. Эта задача существенно упрощается при рассмотрении проекций выпуклых оболочек (или самих классов) на некоторое направление в многомерном пространстве. В более общем случае минимальное расстояние между выпуклыми оболочками и глубину проникновения одной оболочки в другую при их пересечении можно найти, пробуя различные направления в многомерном пространстве и исследуя экстремальные точки проекций классов на эти направления.

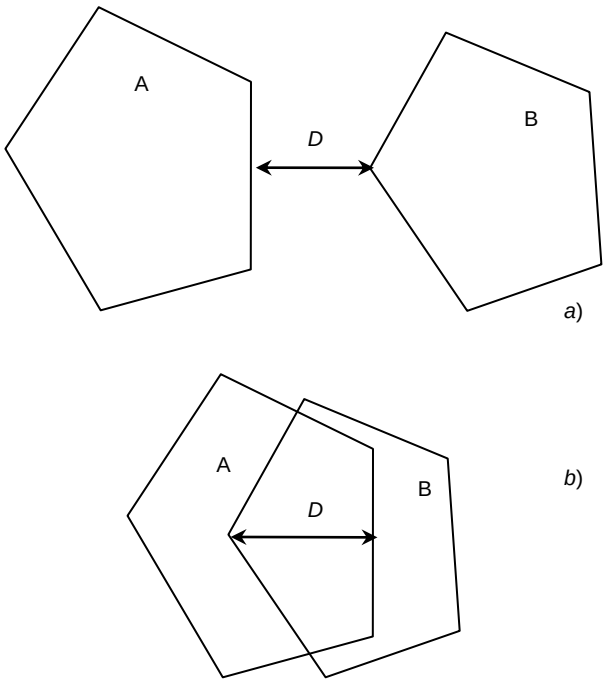

Рис. 1. К определению близости двух выпуклых оболочек друг к другу: a) *D* – минимальное расстояние между A и B; b) *D* – минимальная глубина проникновения между A и B.

Пусть A' и B' проекции классов на некоторое направление в многомерном пространстве. Тогда исключив вырожденный вариант полного включения одного множества в другое меру близости (включая меру пересечения) *D* можно определить в виде следующей процедуры:

$a1 = \min(A')$; $a2 = \max(A')$;
$b1 = \min(B')$; $b2 = \max(B')$;
if $a1 < b1$
    $D = a2 - b1$;
else
    $D = b2 - a1$;
end

В случае пересечения множеств *D* будет иметь отрицательный знак. Для нахождения минимального значения *D* для двух классов нужно найти такое

направление **w** в многомерном пространстве, для которого *D* было бы минимальным.

Многие методы классификации решают проблему неразделимых классов с помощью либо минимизации ошибок классификации, либо с привлечением процедур минимизации таких ошибок. Получаемый при этом весовой вектор, в общем случае, не совпадает с требуемым вектором глубины проникновения для получения визуальной картины на плоскости.

В SVM эта проблема решается с помощью минимизации ошибок классификации, что также не является наилучшим решением с точки зрения критерия минимизации глубины проникновения. При линейно неразделимых классах в методе SVM вместо (1) используется выражение (2).

$$\begin{cases} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \to \min\limits_{\mathbf{w},w_0,\xi_i} \\ y_i\left(\mathbf{w}^T\mathbf{x}_i + w_0\right) \ge 1-\xi_i, \quad i=1,2,...,N, \\ \xi_i \ge 0, \quad i=1,2,...,N \end{cases} \qquad (2)$$

где переменные $\xi_i \ge 0$ отражают величину ошибки на объектах $\mathbf{x}_i$, $i=1,2,...,N$, а коэффициент C - параметр настройки метода, который позволяет регулировать отношение между максимизацией ширины разделяющей полосы и минимизацией суммарной ошибки. Задача решается аналогично задаче для случая линейно разделимых классов.

IV. Использование модифицированных методов SVM

Универсальные способы решения проблемы независимо от условий пересекаемости классов предложены путем трансформации выпуклых оболочек в уменьшенные выпуклые оболочки (RCH - reduced convex hull) [10] и масштабируемые выпуклые оболочки (SCH - scaled convex hull) [11], что сводит задачу к анализу линейно разделимых классов. Нами предложена аналогичная процедура смещенной выпуклой оболочки (OCH – offset convex hull), в результате которой все элементы одного класса смещаются на постоянную величину в направлении вектора разности их центроидов. Далее решается задача с разделенными классами, после чего производится обратное смещение.

V. Экспериментальные исследования

Степень пересекаемости классов после их отображения на плоскость оценивалась числом g членов обучающих выборок обоих классов, попавших в зону пересечения, т.е. $g=(n_1+n_2)/(N_1+N_2)$, 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 $0 < g < 100$ %.

В первом эксперименте по визуализации 4-мерных данных использованы 2 класса ирисов Фишера [12]: виргинского (virginica) и разноцветного (versicolor). Каждый класс состоит из 50 экземпляров, измеренных по 4 признакам: длине и ширине чашелистика; длине и ширине лепестка. Результаты пересекаемости классов после их отображения на плоскость с помощью разных алгоритмов отражены в следующей таблице.

ТАБЛИЦА I ПЕРЕСЕКАЕМОСТЬ КЛАССОВ НА ПЛОСКОСТИ ДЛЯ РАЗНЫХ АЛГОРИТМОВ

Алгоритм	N_1+N_2	n_1	n_2	g %
PCA	50 + 50	2	6	8
FDA	50 + 50	1	2	3
SVM	50 + 50	1	0	1

Эти результаты показывают, что для целей визуализации 2-классовой задачи, заданной в многомерном признаковом пространстве, из исследованных трех методов наилучшим является метод SVM. Он дает минимальное пересечение классов при их отображении на плоскости. Однако этот метод в каждом отдельном случае требует подбора параметров.

Вторая исследованная задача – это диагностика рака молочной железы. Данные взяты из базы Breast Tissue [13]. Они состоят из 106 экземпляров ткани молочной железы измеренных по 9 параметрам импеданса ткани. Данные верифицированы по 6 классам новообразований молочной железы, из которых для наших экспериментов выбраны 2 класса: карцинома молочной железы (злокачественная опухоль) 21 случай и фиброаденома (доброкачественная опухоль) 15 случаев. Исходные данные были пронормированны по среднему значению и дисперсии. Результат применения алгоритма PCA к этим данным показан на рис. 2. Результат визуализации данных с помощью алгоритма SVM, изображенный на рис. 3 показал их полную линейную разделимость.

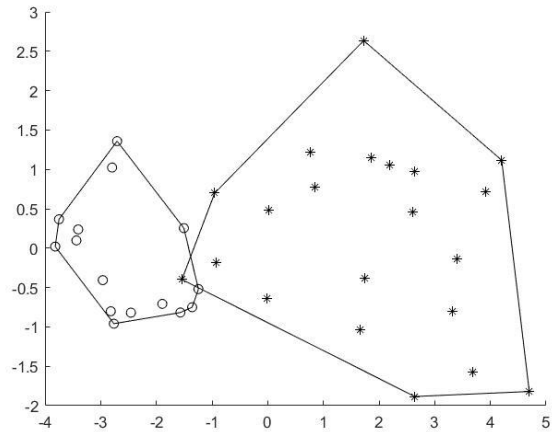

Рис. 2. Отображение классов новообразований молочной железы на плоскости с помощью метода PCA. Слева расположен класс фиброаденомы, справа – класс карциномы. Ось абсцисс – первый весовой вектор, ординат – второй. Видно, что выпуклые оболочки классов пересекаются.

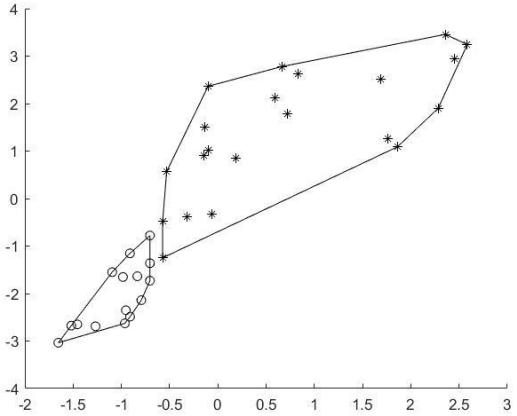

Рис. 3. Отображение классов новообразований молочной железы на плоскости с помощью метода SVM. Взаимное расположение классов и осей то же что и на рис.2. Классы полностью линейно разделимы.

VI. Заключение

Для отображения области пересечения классов из многомерного пространства на плоскость можно использовать критерий близости их выпуклых оболочек *D*. Для непересекающихся классов этот критерий воплощается в минимизацию расстояния между выпуклыми оболочками. Для пересекающихся классов он преобразуется в минимизацию степени их взаимного пересечения *D*. Критерий *D* автоматически выполняется при использовании метода SVM для линейно разделимых классов. При линейно неразделимых классах целесообразно использовать в качестве приближенных решений метод SVM с трансформациями типа RCH, SCH и OCH. Последний из них наиболее прост. При этом вместо SVM допустимо использовать другие алгоритмы ближайшей точки NPP. В общем случае для нахождения оптимальных значений *D* надо использовать поисковые процедуры. Из исследованных трех методов отображения наилучшим оказался метод SVM, который дал минимальное пересечение классов при их отображении на плоскости.

Список литературы

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

[2] 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).

[3] 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).

[4] Maszczyk T. and Duch W. Support Vector Machines for visualization and dimensionality reduction. Lecture Notes in Computer Science, Vol. 5163, 346-356, 2008.

[5] Duda R.O., Hart P.E., Stork D.G.: Pattern Classification. New York: Wiley, 2001. 659 p.

[6] Cortes C. and Vapnik V. N., "Support vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995.

[7] Franc V. and Hlavác V., "An iterative algorithm learning the maximal margin classifier," Pattern Recognit., vol. 36, no. 9, pp. 1985–1996, Sep. 2003.

[8] Mitchell B.F.,. Demyanov V.F, Malozemov V.N., Finding the point of a polyhedron closest to the origin, SIAM J. Control 12 (1974) 19–26.

[9] Weller R., New Geometric Data Structures for Collision Detection and Haptics, Springer Series on Touch and Haptic Systems, DOI 10.1007/978-3-319-01020-5_2, © Springer International Publishing Switzerland 2013.

[10] Mavroforakis M., Sdralis M., Theodoridis S. "A geometric nearest point algorithm for the efficient solution of the SVM classification task," IEEE Transactions on Neural Networks, Vol. 18(5), pp. 1545–1550, 2007.

[11] Zhenbing Liu, J. G. Liu, Chao Pan, and Guoyou Wang. A Novel Geometric Approach to Binary Classification Based on Scaled Convex Hulls, IEEE Transactions on Neural Networks. - 2009. - Vol. 20, No. 7 - pp. 1215–1220.

[12] Iris Data Set. UCI Machine Learning Repository. Available at: https://archive.ics.uci.edu/ml/datasets/iris (accessed 26 April 2018)

[13] Breast Tissue Data Set. UCI Machine Learning Repository. Available at: http://archive.ics.uci.edu/ml/datasets/breast+tissue (accessed 26 April 2018)

[14] Vapnik V. N., Statistical Learning Theory. New York: Wiley, 1998.

[15] Bennett K. P. and Bredensteiner E. J., Duality and geometry in SVM classifiers, in Proc. 17th Int. Conf. Mach. Learn., 2000, pp. 57–64.

[16] Ming C. Lin, Dinesh Manocha, and Young J. Kim, COLLISION AND PROXIMITY QUERIES. In the Handbook of Discrete and Computational Geometry, J.E. Goodman, J. O'Rourke, and C. D. Tóth (editors), 3rd edition, CRC Press, Boca Raton, FL, 2017.