ПРИМЕНЕНИЕ МЕТОДА ОПОРНЫХ ВЕКТОРОВ ДЛЯ КЛАССИФИКАЦИИ ДАННЫХ С ТЕРАГЕРЦОВОГО СПЕКТРОМЕТРА

Ю.К. Изместьева
Научные руководители: профессор, д.ф.м.н. А.В. Шаповалов; к.ф.м.н. А.В. Борисов
Национальный исследовательский Томский политехнический университет,
Россия, г. Томск, пр. Ленина, 30, 634050
E-mail: riitutoriisa@gmail.com

SUPPORT VECTOR MACHINE APPLICATION IN CLASSIFICATION OF TERAHERTZ SPECTROMETER'S DATA

Yu.K. Izmestyeva
Scientific Supervisors: Prof., Dr. A.V. Shapovalov, PhD A.V. Borisov
Tomsk Polytechnic University, Russia, Tomsk, Lenin str., 30, 634050
E-mail: riitutoriisa@gmail.com

Abstract. *Usually data analysis by classic methods is not possible due to the fact that the input data are incomplete or heterogeneous. The solution to this lies in the area of artificial intelligence – specifically, in machine learning discipline. In this article, we consider one of the solutions to the problem of classification by machine learning method – supporting vector machine, by the data taken from the terahertz spectrometer.*

Введение. Область исследования данной научно-исследовательской работы затрагивает такую большую тему, как машинное обучение, в частности – задачу классификации (обучение по прецедентам).

Проблема этой задачи следует прямо из её названия – необходимо определить, к какому классу относится некий элемент или элементы на основе определённых признаков, которыми они обладают. Метод решения, используемый мной в данной работе – это метод опорных векторов (SVM). Этот метод используется, так как имеет ряд преимуществ перед другими методами – он позволяет провести более уверенную классификацию, с его помощью задача решается квадратичным программированием, которое имеет только одно решение, его проще оптимизировать.

Цель работы. Классифицировать данные, полученные с терагерцового спектрометра методом опорных векторов.

Задачи.

- Изучение литературы по областям, затрагиваемым в данной научной работе;
- Приобретение навыков по работе со средой программирования MatLab;
- Интерпретация задачи в MatLab и её решение для имеющихся данных;
- Обсуждение результатов и формирование вывода по проделанной работе.

Материалы и методы исследования. Метод опорных векторов (SVM) – это задача обучения по прецедентам $\langle X, Y, y^*, X^l \rangle$ [1], где X – пространство объектов, Y – множество объектов, $y^* : X \to Y$ – целевая зависимость, $X^l = (x_i, y_i)_{i=1}^l$, $y_i = y^*(x_i)$.

Целью метода является построение алгоритма $a: X \to Y$, аппроксимирующего целевую зависимость на всём пространстве X. Имеем два непересекающихся класса, объекты которых описываются n-мерными вещественными векторами: $X = R^n, Y = -1,1$. Линейный пороговый классификатор имеет вид: $a(x) = \text{sign}\left(\sum_{j=1}^{n} \omega_j x^j - \omega_0\right) = \text{sign}(\omega, x - \omega_0)$, где $\vec{x} = (x^1, \ldots, x^n)$ - признаковое описание объекта; $\vec{\omega}, \omega_0$ – параметры алгоритма; $\omega, x = \omega_0$ – гиперплоскость, разделяющая классы [2].

Решение задачи можно разделить на два случая – когда выборка линейно разделима и не разделима. Особо применяемым при решении проблемы неразделимости является использование ядер и стяжающих пространств, который позволяют осуществить переход от исходного пространства признаковых описаний объектов X к стяжающему пространству H с помощью некоторого преобразования $\psi: X \to H$, в котором выборка может оказаться линейно разделимой.

Некоторые ядра представлены ниже:

- однородный полином $k(x_i, x_j) = (x_i * x_j)^d$,
- неоднородный полином $k(x_i, x_j) = (x_i * x_j + 1)^d$,
- гауссовская функция радиального базиса $k(x_i, x_j) = \exp(-yx_i - x_j^2)$, $y > 0 \text{ или } y = \frac{1}{2\sigma^2}$.

SVM имеет ряд преимуществ перед другими алгоритмами схожего назначения, такими как C4.5, метод k-средних, Apriori, EM-алгоритм. Например такие алгоритмы как C4.5, строящий классификатор в виде дерева решений, или метод k-средних, который создает k -групп из набора объектов таким образом, чтобы члены группы были наиболее однородными, являются более простыми в понимании и интерпретации, но при этом довольны чувствительны к шумам. Также существуют ещё 2 алгоритма – алгоритм Apriori и EM-алгоритм, у которых есть существенный недостаток, выраженный в производительности при большом объёме данных [3].

Используется SVM к данным, получаемым при спектроскопии терагерцовых (ТГц) частотных диапазонов. ТГц излучение применяется очень часто для научных исследований и в прикладных областях: астрофизика, экология, системы связи и др. Одними из самых развивающихся сейчас областей являются медицина и биология. Здесь ТГц спектроскопия позволяет заниматься идентификацией биомолекул, в том числе определение их мутаций; изучением биологических тканей (например, подповерхностных слоев, их диагностика на глубину поражения), обнаружением опухолей, некрозов и других патологических процессов. Спектроскопический анализ выдыхаемого воздуха в ТГц диапазоне может быть эффективным неинвазивным диагностическим средством. Существенный плюс ТГц излучения - оно не является ионизирующим и, следовательно, опасным для биологических объектов, в сравнении с часто используемым рентгеновским [4].

Результаты. Исследуемая выборка представляла собой два набора данных - спектров, снятых с больного и здорового человека терагерцовым спектрометром. Работа с данными осуществлялась в пакете Matlab. Перед проведением классификации методом опорных векторов было необходимо преобразовать входные данные, а именно – интерполировать все исследуемые значения относительно длин волн каждого из них. Для этого задавалась допустимая для всех интерполяционная сетка.

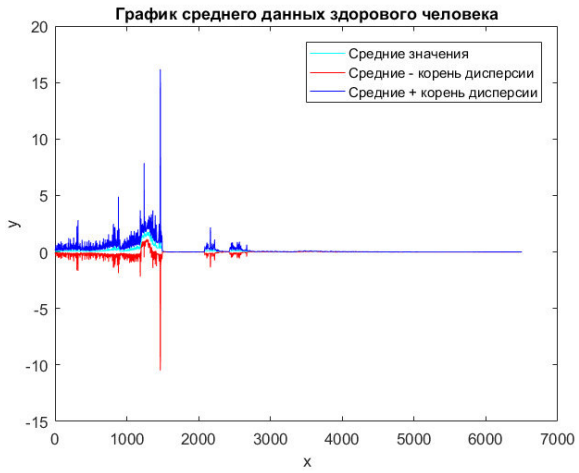

Рис.1. График среднего данных здорового человека

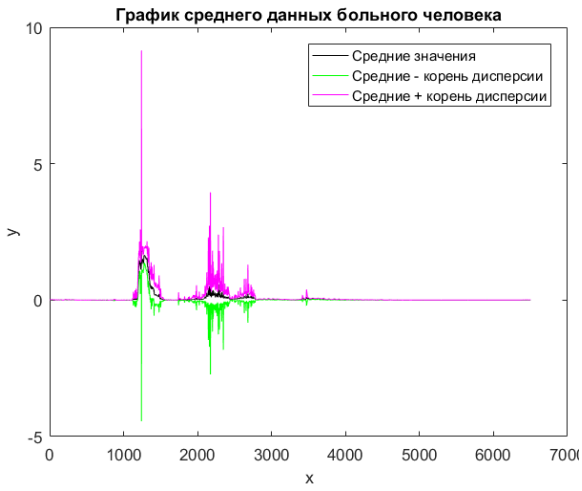

Рис.2. График среднего данных больного человека

На рис. 1, 2 заметны резкие скачки графиков. Они обуславливаются сбоем в самом аппарате, т.е. их можно причислить к приборной погрешности. После интерполяции данных каждый набор спектров разбивался на две части. Первые из них брались для тренировки классификатора, вторые – для непосредственной классификации. Для построения SVM использованы два ядра – радиальную базисную функцию Гаусса и сигмоиду, а затем ядра были исследованы на погрешность в классификации. Погрешность для функции Гаусса составила 9,8%, а погрешность для сигмоиды – 39,22%. Исходя из этих показателей было выбрано первое ядро для дальнейшей работы. Два класса – «больной», «здоровый» обозначены в задаче как «1» и «0» соответственно. Применяя обученный классификатор ко второй части выборок спектров получена итоговую классификацию:

Таблица 1

Значения классов исследуемых спектров

№ спектра	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Класс	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
№ спектра	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**
Класс	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
№ спектра	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**
Класс	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	0

Вывод. В ходе работы проведена классификация спектров, снятых при помощи терагерцового спектрометра, методом опорных векторов. Классификация выявила, что большая часть исследованных данных принадлежит группе «больной».

СПИСОК ЛИТЕРАТУРЫ

1. Вьюгин В.В. Математические основы машинного обучения и прогнозирования. – МЦМНО, 2014. – 304 с.
2. Воронцов К.В. Лекции по методу опорных векторов. – 2007. – 18 с.
3. Wu X., Kumar V. Top 10 algorithms in data mining. – Springer-Verlag London Limited, 2007. – 37 с.
4. Vaks V. High-Precise Spectrometry of the Terahertz Frequency Range: The Methods, Approaches and Applications - Journal of Infrared, Millimeter and Terahertz Waves, 2012, V. 33, N. 1, P. 43-53.