

УДК 681.327

У ШИ (КНР), ЯНЬ ЦЗИНБИНЬ (КНР), Р.М. АЛИЕВ, И.Э. ХЕЙДОРОВ

ГИПЕРСФЕРНАЯ МНОГОКЛАССОВАЯ МАШИНА НА ОПОРНЫХ ВЕКТОРАХ

Support vector machines (SVMs) are learning algorithms derived from statistical learning theory. The SVMs approach was originally developed to solve binary classification problems. In this paper SVMs architectures for multiclass problems are discussed. Furthermore we provide a new algorithm called sphere structured SVMs to solve the multiclass problem. We show the algorithm in detail and analyze its characteristics.

В последнее время при решении задач обработки и классификации аудио- и видеосигналов все чаще используются алгоритмы на основе метода опорных векторов (МОВ) [1, 2]. Сфера применения МОВ на настоящий момент включает актуальные научно-технические области, среди которых текстонезависимая верификация дикторов [3], индексация аудиодокументов [4], биомедицинские приложения [5], семантический поиск [6] и др. Широкая распространенность данного метода обусловлена тем, что машина на опорных векторах представляет собой разделяющий классификатор, простой в своей основе, но чрезвычайно эффективный. Принципиальное его отличие от общепринятых методов классификации, таких как скрытые марковские модели или смешанные гауссовские модели, состоит в том, что этот метод прямо аппроксимирует межклассовые границы вместо моделирования распределения вероятностей по обучающему множеству. Для построения сложных схем классификации обычно применяют многоклассовые машины на опорных векторах, созданные на основе двухклассовых МОВ и использующие следующие алгоритмы:

1) one-against-rest («один против остальных»), для которого характерно формирование отдельного классификатора для каждого класса. Недостатки этого метода: необходимость большого количества выборок каждого класса для обучения классификатора и большее количество вычислений при обучении [1, 5];

2) one-against-one («один против одного»), для которого характерно формирование классификатора для каждой парной комбинации классов. Нужно конструировать $(k-1)k/2$ классификаторов для разных k классов. Недостатком этого метода является огромный объем вычислений [2, 6];

3) алгоритм бинарного дерева – модификация алгоритма one-against-rest [3, 4].

Основным недостатком широко используемых на настоящий момент представленных многоклассовых МОВ является большое количество вычислений и, как следствие, – значительное время для принятия решения. Для преодоления данного недостатка нами разработан новый метод многоклассовой МОВ на основе гиперсферы, который характеризуется меньшим количеством вычислений и возможностью легко добавить новые классы без переобучения всей системы.

Гиперсферная МОВ

Пусть $x_i (x_i \in R^N)$ – набор данных, где $i = 1, \dots, m$, m – число входных образцов. Спроектируем пространство R^N в пространство более высокой размерности F , $R^N \rightarrow F, x_i \rightarrow \varphi(x_i)$, и определим функцию ядра $k(x_i, x_j)$ следующим образом:

$$\forall (x_i, x_j) \in R^N, \quad k(x_i, x_j) = (\varphi(x_i), \varphi(x_j)).$$

Для гиперсферной МОВ сначала необходимо найти гиперсферу с центром c и радиусом R , которая включает в себя все образцы класса. При всех возможных значениях радиуса R нужно добиться минимального предельного расстояния [7]:

$$\min \tau(R, \xi, c) = R^2 + \frac{1}{vm} \sum_{i=1}^m \xi_i,$$

где $\|\varphi(x_i) - c\|^2 \leq R^2 + \xi_i$, $\xi_i \geq 0, i = 1, \dots, m$, v – положительный параметр, регулирующий ошибку классификации; ξ_i – набор дополнительных переменных $\xi_i \geq 0, i = 1, \dots, m$, характеризующих величину ошибки на образцах $x_i \in [x_1 \dots x_m]$.

Предельное расстояние минимизируется под влиянием того, что большинство обучающих векторов x_i расположено в пространстве, ограниченном гиперсферой. Используя функцию Лагранжа, получаем

$$L(R, \xi, c, \alpha, \beta) = R^2 + \frac{1}{vm} \sum_{i=1}^m \xi_i - \sum_{i=1}^m (\alpha_i (R^2 + \xi_i - \|\varphi(x_i) - c\|^2) + \beta_i \xi_i),$$

$$\frac{\partial L}{\partial R} = 2R - 2R \sum_{i=1}^m \alpha_i = 0,$$

$$\frac{\partial L}{\partial \xi_i} = \frac{1}{vm} - \alpha_i - \beta_i = 0,$$

$$\frac{\partial L}{\partial c} = - \sum_{i=1}^m \alpha_i (2\varphi(x_i) - 2c) = 0,$$

где α_i, β_i – коэффициенты Лагранжа, $\alpha_i \geq 0, \beta_i \geq 0$.

Проблема круговой квадратичной оптимизации посредством введения коэффициентов Лагранжа приводит к двойной оптимизационной проблеме:

$$\max W(\alpha_i) = \sum_{i=1}^m \alpha_i K(x_i, x_i) - \sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j),$$

где $0 \leq \alpha_i \leq 1/vm, i = 1, \dots, m; \sum_{i=1}^m \alpha_i = 1$.

Решение этой проблемы можно получить через α_i . Если $\alpha_i \neq 0$, то x_i – опорный вектор. Если $0 < \alpha_k < 1/vm$, то

$$R^2 = \sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j) - 2 \sum_{i=1}^m \alpha_i K(x_i, x_k) + K(x_k, x_k),$$

$$c = \sum_{i=1}^m \alpha_i \varphi(x_i).$$

Для нового вектора z можно получить функцию принятия решений для одного класса МОВ:

$$f(z) = \sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j) - 2 \sum_{i=1}^m \alpha_i K(x_i, z) + K(z, z).$$

Построение гиперсферной многоклассовой машины на опорных векторах

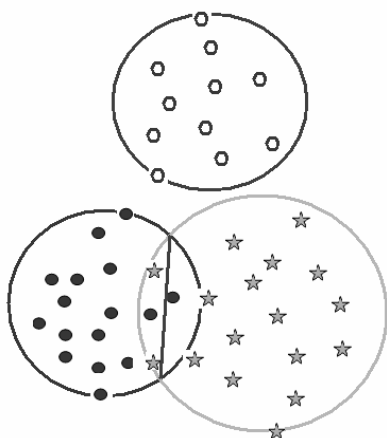
Пусть даны наборы обучающих векторов $\{(x_1, y_1), \dots, (x_l, y_l)\} \subset R^N \times R^N$.

Входное пространство этих векторов разделено на M классов. Каждый класс представляется следующим образом:

$$\{(x_1^{(s)}, y_1^{(s)}), \dots, (x_l^{(s)}, y_l^{(s)}), s = 1, \dots, M\},$$

где l – число векторов для обучения.

Применение многоклассовой МОВ в общем случае сводится к проведению последовательных одно-классовых классификаций для наборов обучающих векторов и получению $R_{(s)}^2$ (квадрат радиуса гиперсферы для класса s). Пример для трех классов приведен на рисунке.



Три гиперсферы для трех классов

Если z – тестируемый вектор, то первоначально нужно вычислить значения функции принятия решений $f_{(s)}(z)$ (квадрат расстояния между z и центром c гиперсферы s класса), а затем произвести сравнение с $R_{(s)}^2$.

Если d – количество гиперсфер, включающих z , то возможно несколько случаев:

1. Если $d = 0$, то вычисляются значения $\gamma_s = f_{(s)}(z) - R_{(s)}^2$, $s = \overline{1, M}$. Минимальное значение γ_s определяет класс-победитель.
2. Если $d = 1$, то z относится к гиперсфере, включающей этот вектор.
3. Если $d > 1$, то z находится на пересечении гиперсфер.

Алгоритм следующий:

Шаг 1: сбор множества номеров I всех гиперсфер, которые содержат z .

Шаг 2: для $\forall i, j \in I$, вычислим проекции $c_i D$ и $c_i C$:

$$c_i D = \sqrt{f(z)_{(i)}} \left(\frac{\|c_j - c_i\|^2 + f(z)_{(i)} - f(z)_{(j)}}{2\sqrt{f(z)_{(i)}} \|c_j - c_i\|} = \frac{\|c_j - c_i\|^2 + f(z)_{(i)} - f(z)_{(j)}}{2\|c_j - c_i\|}, \right.$$

$$c_i C = R_{(i)} \frac{\|c_j - c_i\|^2 + R_{(i)}^2 - R_{(j)}^2}{2R_{(i)} \|c_j - c_i\|} = \frac{\|c_j - c_i\|^2 + R_{(i)}^2 - R_{(j)}^2}{2\|c_j - c_i\|},$$

$$c_i C - c_i D = \frac{f(z)_{(j)} - f(z)_{(i)} + R_{(i)}^2 - R_{(j)}^2}{2\|c_j - c_i\|}.$$

Шаг 3: сравнение значений $c_i D$ и $c_i C$. Если $c_i C - c_i D < 0$, то z относится к гиперсфере c_j , а если $c_i C - c_i D > 0$, то z относится к гиперсфере c_i .

Шаг 4: повторить шаги 2 и 3 для остальных гиперсфер из множества I .

Эксперимент

Для исследования характеристик и границ применимости разработанной многоклассовой гиперсферной МОВ был проведен ряд экспериментов. Предложенная в работе МОВ была реализована в среде Matlab 6.5. Для проведения сравнительных экспериментов с другими МОВ (one-against-rest, one-against-one, бинарное дерево) были взяты стандартные реализации этих алгоритмов, имеющиеся в использованной среде.

Первый эксперимент был посвящен разделению трех классов, выборки которых могут задаваться аналитическими формулами. Были выбраны следующие параметрические функции:

$$\begin{aligned} x_1(t) &= \exp(-0,2t) \cos(2t), & y_1(t) &= \exp(-0,2t) \sin(2t); \\ x_2(t) &= \exp(-0,2(t+1,3)) \cos(2t), & y_2(t) &= \exp(-0,2(t+1,3)) \sin(2t); \\ x_3(t) &= \exp(-0,2(t+2,5)) \cos(2t), & y_3(t) &= \exp(-0,2(t+2,5)) \sin(2t). \end{aligned}$$

Интервал значений времени t составлял от 0 до 10 с шагом 0,02. Количество образцов для обучения и для тестирования – по 300, параметр $\nu = 0,1$. Для гиперсферной многоклассовой МОВ в качестве ядра использовалась экспоненциальная радиальная функция:

$$k(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2).$$

Результаты эксперимента представлены в табл. 1, исходя из которых можно утверждать, что эксперимент подтвердил применимость предложенного метода для случая трехклассовой классификации и определил теоретические значения параметров ядра.

Таблица 1

Полученная точность в зависимости от значения σ^2

σ^2	0,1	0,05	0,03	0,01	0,005
Точность классификации, %	66	77	84	96	100

Второй эксперимент был проведен для двухклассовых наборов реальных биомедицинских данных Breast Cancer и Statlog Heart [8]. В случае Breast Cancer для обучения было использовано 200 выборок и 77 для тестирования. Для данных Statlog Heart – 170 выборок для обучения и 100 для тестирования. В качестве ядра МОВ использовалась экспоненциальная радиальная функция.

Как видно из результатов эксперимента (табл. 2), предложенный алгоритм гиперсферы для двух классов позволяет достичь лучшей либо аналогичной точности по сравнению с обычной МОВ при существенном сокращении временных затрат.

Таблица 2

Сравнение точности двухклассовой классификации и времени вычислений для алгоритма гиперсферы и обычной МОВ

Алгоритм гиперсферы		МОВ-метод	
Точность, %	Время, с	Точность, %	Время, с
Breast Cancer (2-го класса)			
88,3	120,7	85,7	312,1
Statlog Heart (2-го класса)			
83,0	41,9	83,0	50,8

Третий эксперимент был проведен для трехклассовых наборов данных Iris и Wine и шестиклассовой выборки Class [8]. Для данных Iris было использовано 100 выборок для обучения и 50 для тестирования, для Wine – 133 и 45 и для Class – 142 и 72 соответственно. Ядром являлась экспоненциальная радиальная функция. Для данных Iris значение σ^2 было выбрано равным 0,1, для Wine $\sigma^2 = 1,5$, для Class $\sigma^2 = 1,5$. Параметр $\nu = 0,1$.

Как видно из представленных в табл. 3 результатов, использование предложенного алгоритма гиперсферы позволяет добиться примерно аналогичной точности по сравнению с известными алгоритмами при меньших вычислительных затратах, что приводит к уменьшению времени обучения системы и принятия решения в 1,4÷1,9 раза.

Таблица 3

Сравнение точности классификации и времени вычислений для алгоритма гиперсферы и других алгоритмов МОВ

one-against-one		one-against-rest		Бинарное дерево		Алгоритм гиперсферы	
Точность, %	Время, с	Точность, %	Время, с	Точность, %	Время, с	Точность, %	Время, с
Iris (3-го класса)							
96,4	23,4	95,6	23,2	96,1	17,7	96,00	12,25
Wine (3-го класса)							
95,2	58,6	95,4	58,6	95,5	44,8	97,78	23,54
Class (6-го класса)							
79,8	114,3	80,3	103,3	82,3	92,5	80,56	53,04

Таким образом, нами был предложен и исследован метод на основе гиперсферной многоклассовой машины на опорных векторах и показано, что он уменьшает вычислительную сложность классификации, позволяет легко добавлять новые классы, не мешающие прежним и не требующие переобучения всей системы.

1. Bottou L., Cortes C., Denker J. et al. // Proceedings of the International Conference on Pattern Recognition. 1994. Vol. 2. P. 77.
2. Krebel U. Advances in Kernel Methods: Support Vector Learning. London, 1998. P. 255.
3. Zhao Jain, Dong Yuan, Zhao Xianyu et al. // Tsinghua Science and Technology. 2008. Vol. 13. P. 522.
4. Temko A., Nadeu C. // Pattern recognition. 2006. Vol. 39. P. 682.
5. Ладеев Д.Д., Ши У, Хейдоров И.Э. // Proc. of AITTH'2008. Minsk, 2008. P. 212.
6. Ayache S., Quénot G., Gensel J. // Springer. 2007. Vol. 4425. P. 494.
7. Tax D.M., Duin R.P. // Mach. Learn. 2004. Vol. 54. P. 45.
8. <http://archive.ics.uci.edu/ml/datasets.html>

Поступила в редакцию 09.11.08.

У Ши – аспирант кафедры радиофизики. Научный руководитель – И.Э. Хейдоров.

Янь Цзинбинь – аспирант кафедры радиофизики. Научный руководитель – И.Э. Хейдоров.

Роман Мурадович Алиев – студент 5-го курса факультета радиофизики и электроники.

Игорь Эдуардович Хейдоров – кандидат физико-математических наук, доцент кафедры радиофизики.