Московский Физико-Технический Институт (Государственный Университет)

Факультет Управления и Прикладной Математики

Кафедра Интеллектуального Анализа Данных

ДИПЛОМНАЯ РАБОТА БАКАЛАВРА

«Метод опорных признаков в задаче обучения распознаванию объектов двух классов»

Выполнил:

студент 4 курса 774 группы

Савинов Николай Анатольевич

Научный руководитель:

д.т.н., профессор

Моттль Вадим Вячеславович

Заведующий кафедрой Интеллектуального Анализа Данных, член-корреспондент РАН _____ К. В. Рудаков

К защите допускаю

«_____» _____ 2011 г.

К защите рекомендую

«_____» _____ 2011 г.

Москва, 2011

Содержание

Аннотация

В задачах бинарной классификации часто встречаются случаи, когда размерность признакового описания объектов оказывается больше размера обучающей совокупности. В таких ситуациях классический метод SVM склонен к переобучению и оказывается неэффективен, поскольку не производит отбор признаков. В данной работе предлагается параметрическая l_2-l_1 регуляризация для SVM, которая приводит к методу опорных признаков, позволяющему снижать размерность признакового описания. Регуляризатор содержит параметр селективности, определяющий степень отбора признаков. Оптимальное значение селективности выбирается по критерию скользящего контроля Leave One Out. Преимущества метода опорных признаков перед исходным методом опорных векторов продемонстрировано в вычислительных экспериментах на модельных и реальных данных.

1 Введение

Одним из самых популярных современных методов машинного обучения является метод опорных векторов, предложенный В. Н. Вапником [1]. В англоязычной литературе он известен под названием SVM (Support Vector Machine).

Этот метод обладает несколькими замечательными свойствами. Во-первых, его обучение может осуществляться с помощью эффективного алгоритма Sequential Minimal Optimization, предложенного в [2], который позволяет работать с десятками тысяч объектов в обучающей совокупности. Во-вторых, результат обучения обладает свойством разреженности - лишь небольшая доля объектов обучения влияет на положение разделяющей гиперплоскости, определяемой линейным решающим правилом. В-третьих, для линейно разделимой выборки метод максимизирует ширину разделяющей полосы между классами, что улучшает обобщающую способность.

Однако в тех ситуациях, когда размерность признакового описания объектов оказывается больше размера обучающей совокупности, SVM уступает другим регуляризованным классификаторам [3]. Причиной является то, что l_2-регуляризатор, используемый в SVM для улучшения обобщающей способности, не отбирает признаки. В таких ситуациях оценки обобщающей способности Вапника-Червоненкиса [1] указывают на переобучение. Следовательно, для задач такого типа необходима дополнительная регуляризация.

В данной работе предлагается параметрическая l_2-l_1 регуляризация, которая заключается во введении в функцию потерь SVM комбинации норм l_2 и l_1 направляющего вектора разделяющей гиперплоскости. l_1 норма была выбрана потому, что она отвечает лапласовской регуляризации, успешно примененной в [4] для отбора признаков линейной регрессии. Предложенный метод назван методом опорных признаков.

В предлагаемом регуляризаторе присутствует параметр селективности, позволяющий регулировать степень отбора. При нулевой селективности метод соответствует классическому SVM, при неограниченном приближении селективности к 1 метод отбрасывает все признаки. Подбор параметра селективности предлагается осуществлять по Cross Validation.

Обучение по предложенному критерию реализуется решением двойственной вогнутой задачи, которая не является квадратичной, но может быть сведена к задаче квадратичного программирования за счет введения дополнительных переменных оптимизации. Такой метод решения накладывает определенные ограничения на размерность задачи, поскольку эффективный алгоритм SMO напрямую неприменим в данном случае. Поэтому в работе предлагается идея обобщения алгоритма SMO для неквадратичной целевой функции, что позволяет решать двойственную задачу без введения дополнительных переменных.

Экспериментальное исследование на модельных и реальных данных, проведенное в данной работе, доказывает эффективность метода опорных признаков по сравнению с классическим методом опорных векторов.

1.1 Метод опорных векторов

Рассмотрим постановку задачи распознавания объектов двух классов.

Задана обучающая выборка $(\mathbf{x}_j, y_j),\ j = 1, ..., N$, где $\mathbf{x_j} \in \mathbb{R}^n$ — признаковые описания объектов, $y_j \in \{-1, 1\}$ — ответы.

Рассматривается класс линейных решающих правил: $y = sign(\sum\limits_{i=1}^{n} a_i x_i + b)$.

Тогда разделяющая гиперплоскость в пространсве признаковых описаний объектов описывается уравнением $\sum\limits_{i=1}^{n} a_i x_i + b = 0$. Линейный классификатор относит новый объект к одному из двух классов в зависимости от того, с какой стороны гиперплоскости оказалась точка, отвечающая его признаковому описанию.

Сначала рассмотрим линейно разделимую выборку. В этом случае можно провести бесконечное множество разделяющих гиперплоскостей. Чтобы избежать переобучения (ясно, что произвольная гиперплоскость скорее всего не является оптимальной), В. Н. Вапник предложил максимизировать ширину разделяющей полосы между классами, то есть чтобы разделяющая гиперплоскость максимально далеко отстояла от ближайших к ней точек обоих классов. Это требование дает критерий обучения, позволяющий оценить параметры $\mathbf{a}$ и b:

$$\begin{cases} \mathbf{a}^T\mathbf{a} \to \min, \\ y_i(\mathbf{a}^T\mathbf{x}_j + b) \geq 1,\ j = 1, ..., N. \end{cases} \tag{1}$$

В критерии появляется регуляризатор $\mathbf{a}^T\mathbf{a}$, представляющий собой квадрат нормы l_2 направляющего вектора разделяющей гиперплоскости.

Если выборка линейно неразделима, оптимизационный критерий (1) не имеет решения. Поэтому требуется его обобщение. Вводятся дополнительные переменные $\delta_j \geq 0, \quad \text{j} = 1, ..., \text{N}$, отвечающие величинам ошибок на объектах выборки, причем ошибкой считается не только попадание по другую сторону гиперплоскости, но и в разделяющую полосу. Обобщенный критерий имеет следующий вид:

$$\begin{cases} \frac{1}{2}\mathbf{a}^T\mathbf{a} + c\sum\limits_{j=1}^{N}\delta_j \to \min, \\ y_i(\mathbf{a}^T\mathbf{x}_j + b) \geq 1 - \delta_j, \\ \delta_j \geq 0, \ j = 1, ..., N. \end{cases} \tag{2}$$

Двойственная к данной задача имеет вид:

$$\begin{cases} \sum\limits_{j=1}^{N}\lambda_j - \frac{1}{2}\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j \to max(\lambda_j), \\ 0 \leq \lambda_j \leq c, \ j = 1, ..., N, \\ \sum\limits_{j=1}^{N} y_j\lambda_j = 0. \end{cases} \tag{3}$$

Если найдено решение двойственной задачи, решение прямой задачи находится из условий Каруша-Куна-Таккера:

- $\mathbf{a} = \sum\limits_{j=1}^{N}\lambda_j y_j \mathbf{x}_j$
- $b = median\{y_j - \mathbf{a}^T\mathbf{x}_j, \ \lambda_j > 0, \ j = 1, ..., N\}$

Заметим, что направляющий вектор $\mathbf{a}$ разделяющей гиперплоскости является линейной комбинацией признаковых описаний объектов обучения. При этом на положение разделяющей гиперплоскости влияют только те объекты, для которых двойственные переменные $\lambda_j > 0$ (их называют опорными объектами).

Для решения двойственной задачи John Platt предложил алгоритм SMO, который согласно экспериментальному исследованию [2] выполняется за $O(N^2)$ действий. Это существенно лучше стандартных методов решения задачи квадратичного программирования, которые имеют асимптотическую сложность $O(N^3)$. Алгоритм основан на следующих идеях:

1. Решение двойственной задачи обычно оказывается разреженным, то есть только небольшая доля двойственных переменных отлична от 0. Нужно лишь определить, какие объекты являются опорными и искать значения двойственных переменных именно для них.

2. Если оптимизировать двойственные переменные парами, фиксируя остальные, можно найти решение аналитически. Это позволяет избежать применения сложных оптимизационных процедур, подверженных машинным ошибкам округления.

Таким образом, весь процесс обучения SVM выполняется достаточно быстро и позволяет работать с десятками тысяч объектов в обучении. Чем больше обучающая выборка, тем больше будет точность на контрольной выборке, а благодаря эффективному алгоритму SMO с этой точки зрения нет границ для наращивания точности. Однако в противоположной ситуации, когда объектов в обучающей совокупности слишком мало по сравнению с числом признаков, SVM оказывается неэффективен. В таких случаях необходима регуляризация, позволяющая осуществлять отбор признаков и снижать размерность задачи.

1.2 Введение l_2-l_1 регуляризации

В данной работе предлагается ввести следующий регуляризатор, который заменит слагаемое $\frac{1}{2}\mathbf{a}^T\mathbf{a}$ в целевой функции классического критерия метода опорных векторов (2):

$$\sum_{i=1}^{n}((1-\mu)a_i^2+\mu|a_i|), \text{где } \mu \in [0,1) \text{ — параметр, позволяющий регулировать селективность метода.} \tag{4}$$

Кроме того, что использование l_1-нормы для регуляризации оказалось эффективно в известном регрессионном методе LASSO [4], можно привести следующие неформальные соображения в пользу применения (4) для отбора признаков. У большинства задач анализа данных есть вероятностная байесовская постановка, предполагающая наличие априорного распределения на множестве параметров алгоритма. Регуляризатор l_2 соответствует нормальному распределению, l_1 - распределению Лапласа.

Если построить график распределения для предлагаемого l_2-l_1 регуляризатора, он будет иметь следующий вид (в двухмерном случае, при значении селективности 0.5):

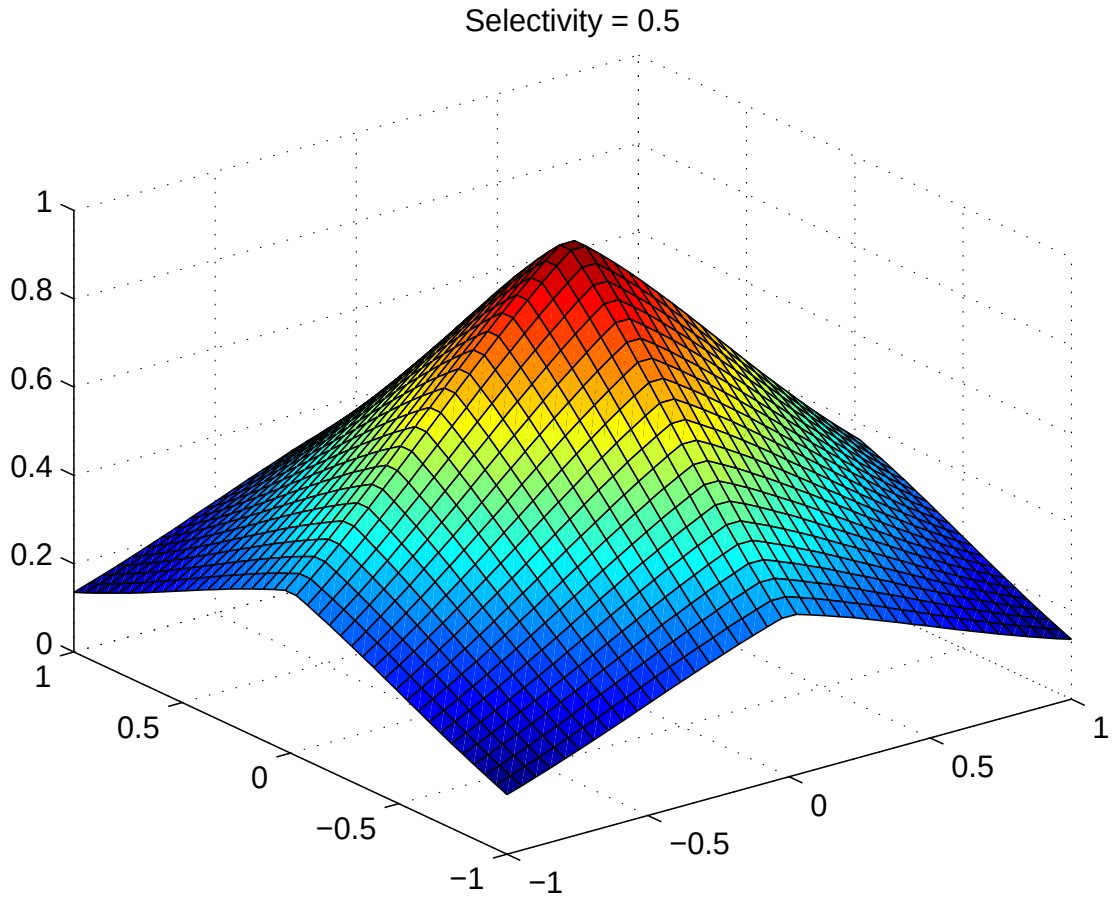

Рис. 1: Априорное распределение компонент a_1, a_2

Следует обратить внимание на наличие острых "ребер"вдоль осей графика. Исходя из байесовкой постановки, если точка, соответствующая набору параметров классификатора, находится близко к одной из осей, но не на самой оси, всегда выгодно будет сместиться на саму ось - это приведет к резкому росту апостериорной вероятности реализации значений параметров (то есть оптимизационный критерий выберет именно решение на оси). А такое смещение будет означать отбор признаков, поскольку значение соответсвующего коэффициента a_i становится равным 0. При этом параметр селективности позволяет регулировать остроту "ребер"распределения, тем самым делая отбор более или менее интенсивным.

Поскольку при отборе останутся лишь некоторые признаки, метод был назван методом опорных признаков.

2 Метод опорных признаков

2.1 Прямая задача

Обучение в предложенном методе сводится к решению следующей выпуклой задачи оптимизации:

$$\begin{cases} \sum\limits_{i=1}^{n}((1-\mu)a_i^2+\mu|a_i|)+c\sum\limits_{j=1}^{N}\delta_j \to \min(\mathbf{a},b,\delta_j), \\ y_j(\mathbf{a}^T\mathbf{x}_j+b)\geq 1-\delta_j,\ \ j=1,...,N, \\ \delta_j\geq 0,\ \ j=1,...,N. \end{cases} \tag{5}$$

Про классический критерий обучения SVM известно, что решение двойственной задачи обладает свойством разреженности - лишь некоторые переменные в точке оптимума отличны от 0. А вычислительное решение разреженной задачи проще, чем задачи, этим свойством не обладающей. Поэтому предлагается и в данном критерии перейти к двойственной задаче, рассчитывая на разреженность ее решения.

2.2 Двойственная задача

Составим функцию Лагранжа для прямой задачи (5). Она будет регулярной, поскольку выполнены условия Слейтера, изложенные в [5]. А именно, в допустимом множестве, определяемом ограничениями, существует внутренняя точка $a_1 = ... = a_n = 0,\ b = 0,\ \delta_1 = ... = \delta_N = 2$. Таким образом, коэффициент при целевой функции можно положить равным 1:

$$L=\sum_{i=1}^{N}\left[(1-\mu)a_i^2+\mu|a_i|)\right]+c\sum_{j=1}^{N}\delta_j-\sum_{j=1}^{N}\pi_j\delta_j-\sum_{j=1}^{N}\lambda_j\left[y_j\left(\sum_{i=1}^{n}a_ix_{ij}+b\right)-1+\delta_j\right]=$$

$$=\sum_{i=1}^{N}\left[(1-\mu)a_i^2+\mu|a_i|-a_i\left(\sum_{j=1}^{N}\lambda_jy_jx_{ij}\right)\right]+\sum\left(c-\lambda_j-\pi_j\right)\delta_j-$$

$$-\left(\sum_{j=1}^{N}\lambda_jy_j\right)b+\sum_{j=1}^{N}\lambda_j\to\begin{cases}\inf(a_i,b,\delta_j)\\ \max(\lambda_j\geq 0,\pi_j\geq 0)\end{cases}$$

Введем обозначения:

$$w_i = \sum_{j=1}^{N} \lambda_j y_j x_{ij} \tag{6}$$

$$L_i = (1-\mu)a_i + \mu|a_i| - w_i a_i \tag{7}$$

Тогда:

$$L = \sum_{i=1}^{n} L_i + \sum (c - \lambda_j - \pi_j)\,\delta_j - \left(\sum_{j=1}^{N} \lambda_j y_j\right) b + \sum_{j=1}^{N} \lambda_j \to \begin{cases} \inf(a_i, b, \delta_j) \\ \max(\lambda_j \geq 0, \pi_j \geq 0) \end{cases}$$

Сначала минимизируем по a_i:

$$\widehat{a_i} = \operatorname*{argmin}_{a_i} L_i$$

Заметим, что L_i - выпуклая по a_i функция. По [5], необходимым и достаточным условием глобального минимума выпуклой функции является: $0 \in \partial L_i(\widehat{a_i})$, где ∂L_i - субдифференциал. Рассмотрим различные случаи:

1. $a_i > 0$:

   $((1-\mu)a_i^2 + \mu a_i - w_i a_i)' = 2(1-\mu)a_i + (\mu - w_i) = 0$

   Отсюда: $\widehat{a_i} = \dfrac{w_i - \mu}{2(1-\mu)}$ — является решением при $w_i > \mu$.

2. $a_i < 0$:

   $((1-\mu)a_i^2 - \mu a_i - w_i a_i)' = 2(1-\mu)a_i + (\mu + w_i) = 0$

   Отсюда: $\widehat{a_i} = \dfrac{w_i + \mu}{2(1-\mu)}$ — является решением при $w_i < \mu$.

3. $a_i = 0$:

   $0 \in \partial((1-\mu)a_i^2 + \mu|a_i| - w_i a_i) = 2(1-\mu)a_i + B_\mu(0) - w_i = B_\mu(0) - w_i,$

   где $B_\mu(0)$ - замкнутый шар радиуса μ с центром в 0.

   Отсюда: $\widehat{a_i} = 0$ — является решением при $w_i \in B_\mu(0)$.

В итоге получаем формулу, по которой направляющий вектор $\mathbf{a}$ выражается через двойственные переменные λ_j (здесь подставлено выражение 6):

$$\widehat{a_i} = \begin{cases} \frac{\sum\limits_{j=1}^{N} \lambda_j y_j x_{ij} - \mu}{2(1-\mu)}, \ \sum\limits_{j=1}^{N} \lambda_j y_j x_{ij} > \mu, \\ \frac{\sum\limits_{j=1}^{N} \lambda_j y_j x_{ij} + \mu}{2(1-\mu)}, \ \sum\limits_{j=1}^{N} \lambda_j y_j x_{ij} < -\mu, \\ 0, \ \sum\limits_{j=1}^{N} \lambda_j y_j x_{ij} \in B_\mu(0). \end{cases} \tag{8}$$

Теперь найдем значение $\widehat{L_i} = L_i(\widehat{a_i})$.

Пусть $|w_i| > \mu$. В этом случае:

$$\widehat{a_i} = \frac{|w_i| - \mu}{|w_i|} w_i, \ |\widehat{a_i}| = \frac{|w_i| - \mu}{2(1-\mu)}.$$

Тогда:

$$\begin{aligned} &\widehat{L_i} = (1-\mu)\widehat{a_i}^2 + \mu|\widehat{a_i}| - w_i\widehat{a_i} = \\ &= (1-\mu)\frac{(|w_i| - \mu)^2}{4(1-\mu)^2} + \mu\frac{|w_i| - \mu}{2(1-\mu)} - \frac{|w_i| - \mu}{2(1-\mu)}|w_i| = \\ &= \frac{1}{4(1-\mu)}\left[(|w_i| - \mu)^2 + 2\mu(|w_i| - \mu) - 2(|w_i| - \mu)|w_i|\right] = \\ &= \frac{1}{4(1-\mu)}\left(-|w_i|^2 + 2\mu|w_i| - \mu^2\right) = -\frac{1}{4(1-\mu)}\left(|w_i| - \mu\right)^2. \end{aligned}$$

Если $|w_i| \leq \mu$ (то есть $w_i \in B_\mu(0)$), то:

$$\widehat{a_i} = 0 \Rightarrow \widehat{L_i} = 0.$$

В итоге:

$$\begin{aligned} &\widehat{L_i} = \min_{a_i} L_i = \begin{cases} -\frac{1}{4(1-\mu)}\left(|w_i| - \mu\right)^2, \ |w_i| > \mu \\ 0, |w_i| \leq \mu \end{cases} = \\ &= -\frac{1}{4(1-\mu)}\left(\min\{0, \mu - |w_i|\}\right)^2 = \\ &= -\frac{1}{4(1-\mu)}\left(\min\{\mu + w_i, 0, \mu - w_i\}\right)^2. \end{aligned}$$

В соответствии с [5], двойственная задача рассматривается на том множестве двойственных переменных Y, для которого $\inf\limits_{(\lambda_j, \pi_j) \in Y} L > -\infty$. Отсюда вытекают огра-

ничения на двойственные переменные:

$$\begin{cases} c - \lambda_j - \pi_j = 0, \\ \sum\limits_{j=1}^{N} y_j \lambda_j = 0, \\ \pi_j \geq 0, \\ \lambda_j \geq 0, \ j = 1, ..., N. \end{cases} \tag{9}$$

На этом множестве целевая функция двойственной задачи имеет вид:

$$W(\lambda) = \min_{a_i, b, \delta_j} L = \sum_{j=1}^{N} \lambda_j + \sum_{i=1}^{n} \widehat{L}_i.$$

Исключив из рассмотрения переменные π_j и подставив выражение (6), получим итоговый вид двойственной задачи:

$$\begin{cases} W(\lambda) = \sum\limits_{j=1}^{N} \lambda_j - \frac{1}{4(1-\mu)} \sum\limits_{i=1}^{n} \left(\min\{\mu + \sum\limits_{j=1}^{N} \lambda_j y_j x_{ij}, 0, \mu - \sum\limits_{j=1}^{N} \lambda_j y_j x_{ij}\} \right)^2 \to \max\limits_{\lambda}, \\ 0 \leq \lambda_j \leq c, \\ \sum\limits_{j=1}^{N} y_j \lambda_j = 0, \ j = 1, ..., N. \end{cases} \tag{10}$$

Заметим, что решение двойственной задачи существует, поскольку целевая функция непрерывна, а допустимое множество является компактом.

2.3 Решение прямой задачи

Пусть решение двойственной задачи λ^* найдено. Покажем, что в этом случае решение прямой задачи аналитически выражается через решение двойственной.

Поскольку прямая задача выпукла и для нее выполнено условие Слейтера, справедлива теорема Куна-Таккера в форме двойственности из [5]. Она дает необходимые и достаточные условия оптимума для прямой задачи. Запишем условия дополняющей нежесткости, являющиеся необходимыми, а также одно из ограничений (9) на допустимое множество двойственной задачи:

$$\begin{cases} \lambda_j (y_j (\mathbf{a}^T \mathbf{x}_j + b) - 1 + \delta_j) = 0, \\ \delta_j \pi_j = 0, \\ \lambda_j + \pi_j = c, \ j = 1, ..., N. \end{cases}$$

Далее:

$$\{0 < \lambda_j < c\} \Rightarrow \{\delta_j = 0\} \Rightarrow \{y_j(\mathbf{a}^T\mathbf{x}_j + b) = 1\} \Rightarrow \{\mathbf{a}^T\mathbf{x}_j + b = y_j\} \Rightarrow \{b = y_j - \mathbf{a}^T\mathbf{x}_j\}$$

Выполняя суммирования по всем $j:\ 0 < \lambda_j < c$ получаем:

$$b = \frac{\sum\limits_{0<\lambda_j<c} \lambda_j y_j - \sum\limits_{0<\lambda_j<c} \lambda_j \mathbf{a}^T\mathbf{x}_j}{\sum\limits_{0<\lambda_j<c} \lambda_j}$$

Используя ограничение $\sum\limits_{j=1}^{N} \lambda_j y_j = 0$ двойственной задачи, можно преобразовать это выражение к виду:

$$b = -\frac{\sum\limits_{0<\lambda_j<c} \lambda_j \mathbf{a}^T\mathbf{x}_j + c\left(\sum\limits_{\lambda_j=c} y_j\right)}{\sum\limits_{0<\lambda_j<c} \lambda_j} \tag{11}$$

Таким образом, выражения (8) и (11) определяют параметры разделяющей гиперплоскости $\mathbf{a}$, b через решение двойственной задачи λ^*. Поскольку значения этих параметров определяются единственным образом, их допустимость проверять не нужно (доказано в [5]). Тогда по теореме Куна-Таккера в форме двойственности $\mathbf{a}$, b, определяемые выражениями (8), (11), являются оптимальными.

Итоговый вид решения прямой задачи, выраженного через решение двойственной:

$$\begin{cases} \widehat{a_i} = \begin{cases} \frac{\sum\limits_{j=1}^{N} \lambda_j y_j x_{ij} - \mu}{2(1-\mu)}, & \sum\limits_{j=1}^{N} \lambda_j y_j x_{ij} > \mu, \\ \frac{\sum\limits_{j=1}^{N} \lambda_j y_j x_{ij} + \mu}{2(1-\mu)}, & \sum\limits_{j=1}^{N} \lambda_j y_j x_{ij} < -\mu, \\ 0, & \sum\limits_{j=1}^{N} \lambda_j y_j x_{ij} \in B_\mu(0), \end{cases} \\ b = -\frac{\sum\limits_{0<\lambda_j<c} \lambda_j \mathbf{a}^T\mathbf{x}_j + c\left(\sum\limits_{\lambda_j=c} y_j\right)}{\sum\limits_{0<\lambda_j<c} \lambda_j}. \end{cases} \tag{12}$$

2.4 Способы решения двойственной задачи

Двойственная задача (10) является вогнутой, но не является квадратичной, как в классическом методе опорных векторов. Поэтому для ее решения нельзя напрямую

использовать методы, разработанные для SVM. Предлагается два способа решения данной проблемы:

1. Сведение к задачи к квадратичной путем введения дополнительных переменных.

2. Решение двойственной задачи напрямую с помощью обобщения метода Sequential Minimal Optimization на данный неквадратичный случай.

Первый способ позволяет использовать стандартные методы численного решения задач квадратичного программирования, такие как метод проекций сопряженных градиентов и метод внутренней точки [6]. Второй способ позволяет избежать увеличения размерности задачи, к которому приводит первый.

2.5 Квадратичная форма двойственной задачи

Рассмотрим двойственную задачу (10). Ее можно записать в следующем виде, используя выражение (6):

$$\begin{cases} W(\lambda) = \sum\limits_{j=1}^{N} \lambda_j - \frac{1}{4(1-\mu)} \sum\limits_{i=1}^{n} (\min\{0, \mu - |w_i|\})^2 \to \max\limits_{\lambda}, \\ 0 \le \lambda_j \le c, \\ \sum\limits_{j=1}^{N} y_j \lambda_j = 0, \ j = 1, ..., N. \end{cases}$$

Заметим: $(\min\{0, \mu - |w_i|\})^2 = (\max\{0, |w_i| - \mu\})^2$.

Введем дополнительные переменные $\gamma_i = \max\{0, |w_i| - \mu\}$.

Тогда задача сведется к следующей:

$$\begin{cases} W(\lambda) = \sum\limits_{j=1}^{N} \lambda_j - \frac{1}{4(1-\mu)} \sum\limits_{i=1}^{n} \gamma_i^2 \to \max\limits_{\lambda}, \\ \gamma_i \ge |w_i| - \mu = |\sum\limits_{j=1}^{N} \lambda_j y_j x_{ij}| - \mu, \ i = 1, ..., n, \\ 0 \le \lambda_j \le c, \\ \sum\limits_{j=1}^{N} y_j \lambda_j = 0, \ j = 1, ..., N. \end{cases} \tag{13}$$

Эта задача является квадратичной, однако размерность увеличилась с N до $(N + n)$ переменных. Задачи квадратичного программирования в общем случае решаются за $O(dim^3)$ действий, где dim — число переменных. Предложенный метод

опорных признаков предназначен для использования в задачах, где $n \geq N$. В таком случае, введение дополнительных переменных увеличит размерность как минимум в 2 раза, а время настройки классификатора увеличится как минимум в 8 раз по сравнению с классическим методом опорных векторов (при условии, что для решения двойственной задачи в нем используются обычные алгоритмы квадратичного программирования, а не эффективный метод SMO).

Следует отметить, что для данной квадратичной задачи метод SMO напрямую неприменим, поскольку появляются дополнительные переменные γ_i, изменяющие структуру оптимизационной задачи. Этого недостатка лишена идея, изложенная в следующем разделе.

2.6 Обобщение метода Sequential Minimal Optimization для решения неквадратичной двойственной задачи

Для использования метода SMO достаточно решить двойственную задачу (10) при 2 свободных переменных λ_{k_1}, λ_{k_2} и фиксированных остальных.

Начнем с вычисления субдифференциала целевой функции, которая имеет вид:

$$W(\lambda) = \sum_{j=1}^{N} \lambda_j - \frac{1}{4(1-\mu)} \sum_{i=1}^{n} [min\{\mu + X_{y,i}\lambda, 0, \mu - X_{y,i}\lambda\}]^2,$$

где $X_{y,i}$ — вектор-строка матрицы $(y_j x_{ij})$, λ — вектор-столбец двойственных переменных. Используя свойства субдифференциала сложной функции:

$$\partial_\lambda W = \mathbf{1} - \frac{1}{2(1-\mu)} \sum_{i=1}^{n} min\{\mu + X_{y,i}\lambda, 0, \mu - X_{y,i}\lambda\} \partial_\lambda (min\{\mu + X_{y,i}\lambda, 0, \mu - X_{y,i}\lambda\}) =$$

$$= \mathbf{1} - \frac{1}{2(1-\mu)} \left(\sum_{i:\, X_{y,i}\lambda < -\mu} (\mu + X_{y,i}\lambda) X_{y,i}^T - \sum_{i:\, X_{y,i}\lambda > \mu} (\mu - X_{y,i}\lambda) X_{y,i}^T \right).$$

Полученный субдифференциал состоит ровно из одного элемента, поэтому функция дифференцируема: $\frac{\partial W}{\partial \lambda} = \partial_\lambda W$. Более того, имеет место непрерывная дифференцируемость.

Пусть для оптимизации выбрана пара двойственных переменных $\lambda_{k_1}, \lambda_{k_2}$, остальные считаем фиксированными. Обозначим: $s = y_{k_1} y_{k_2}$. Тогда из исходной задачи (10) получаем ограничение-равенство:

$$\lambda_{k_2} + s\lambda_{k_1} = \lambda_{k_2}^0 + s\lambda_{k_1}^0 = \gamma, \text{ где индекс 0 означает значение до оптимизации.} \quad (14)$$

Можно выразить: $\lambda_{k_2} = \gamma - s\lambda_{k_1}$. При этом задача двухмерной оптимизации становится одномерной.

Найдем безусловный максимум полученной одномерной функции. Используя правила дифференцирования сложной функции:

$$\frac{dW}{d\lambda_{k_1}} = \frac{\partial W}{\partial \lambda}\frac{\partial \lambda}{\partial \lambda_{k_1}} = (1-s) -$$

$$- \frac{y_{k_1}}{2(1-\mu)} \left(\sum_{i:\, X_{y,i}\lambda < -\mu} (\mu + X_{y,i}\lambda)(x_{ik_1} - x_{ik_2}) - \sum_{i:\, X_{y,i}\lambda > \mu} (\mu - X_{y,i}\lambda)(x_{ik_1} - x_{ik_2}) \right) \Bigg|_{\lambda_{k_2} = \gamma - s\lambda_{k1}}$$

Производная является монотонной кусочно-линейной функцией, состоящей как максимум из $(2n+1)$ участка. Таким образом, существует как максимум $2n$ точек излома. Для нахождения линейного участка, на котором производная обращается в 0, можно использовать метод дихотомии. Зная значения производной на концах этого участка, можно найти точку 0 производной. После этого полученная точка аналитически проекцируется на множество, определяемое ограничениями-неравенствами, и решение двухмерной задачи получено.

Таким образом, алгоритм SMO можно применять для решения данной неквадратичной задачи, если проводить двухмерную оптимизацию по указанному способу.

3 Вычислительные эксперименты

3.1 Модельные данные: 2 гиперкуба

- $n = 50$ — число признаков, $N_1 = 25$ — число объектов первого класса, $N_2 = 25$ — число объектов второго класса.
- Данные генерировались с равномерным распределением внутри двух гиперкубов, касающихся по одной из граней. Нормаль к этой грани направлена по вектору $(5, 4, 3, 2, 1, 0, ..., 0)$.
- Таким образом, число реальных признаков, по которым выборка разделима, равно 5.

Обобщающая способность контролировалась двумя способами:

- На контрольной выборке ($N = 100000$ объектов), которая генерировалась тем же способом, что и обучающая.
- С помощью Cross Validation (а именно, вычислялась доля ошибок классификации на Leave One Out).

3.2 Результаты на модельных данных

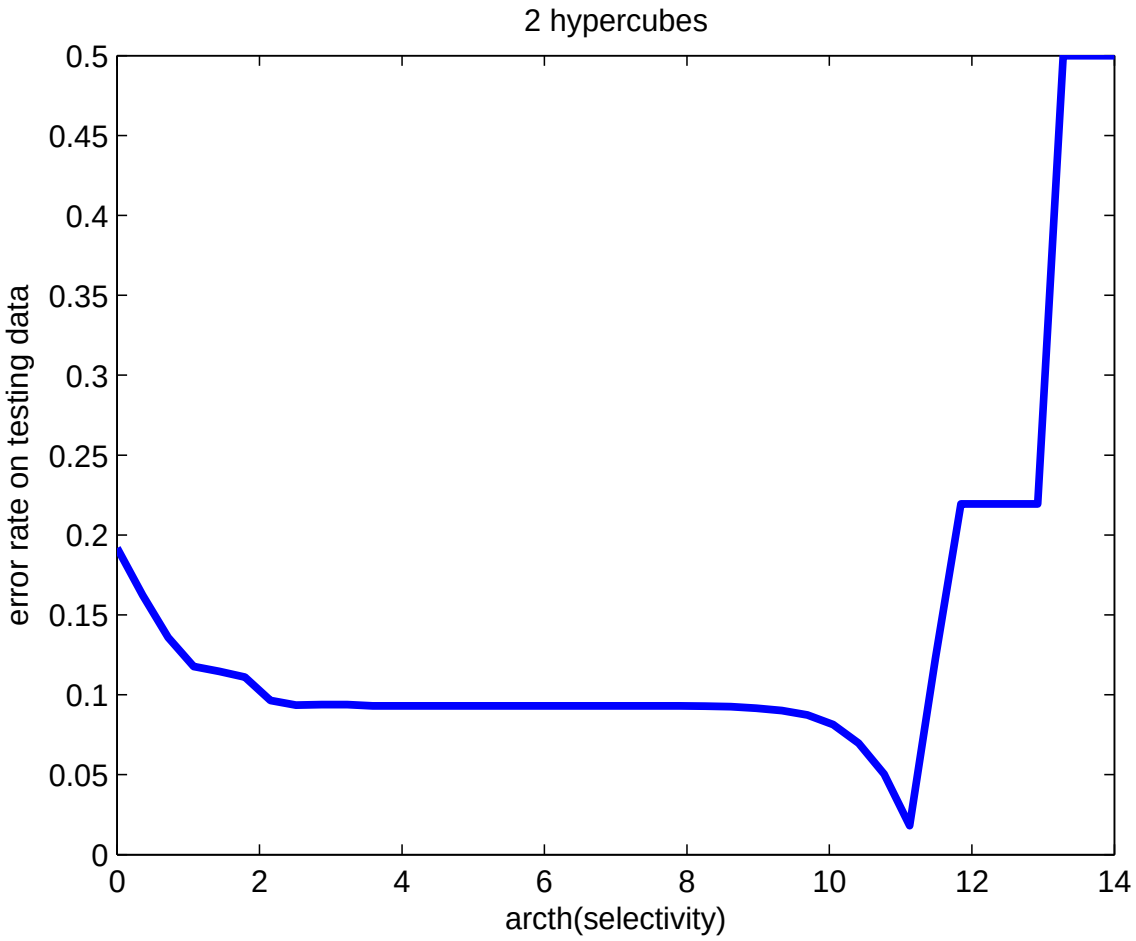

Рис. 2: Зависимость доли ошибок на контрольной выборке от значения параметра селективности

- В точке минимума ошибки алгоритм выделяет 4 признака из 5 правильных.
- Использование метода опорных признаков позволяет уменьшить долю ошибок на контрольной выборке на 17 % по сравнению с обычным SVM (который является частным случаем предложенного метода при значении селективности $\mu = 0$).
- Для целей масштабирования по оси X отложена не селективность μ, а величина $arcth(\mu)$.

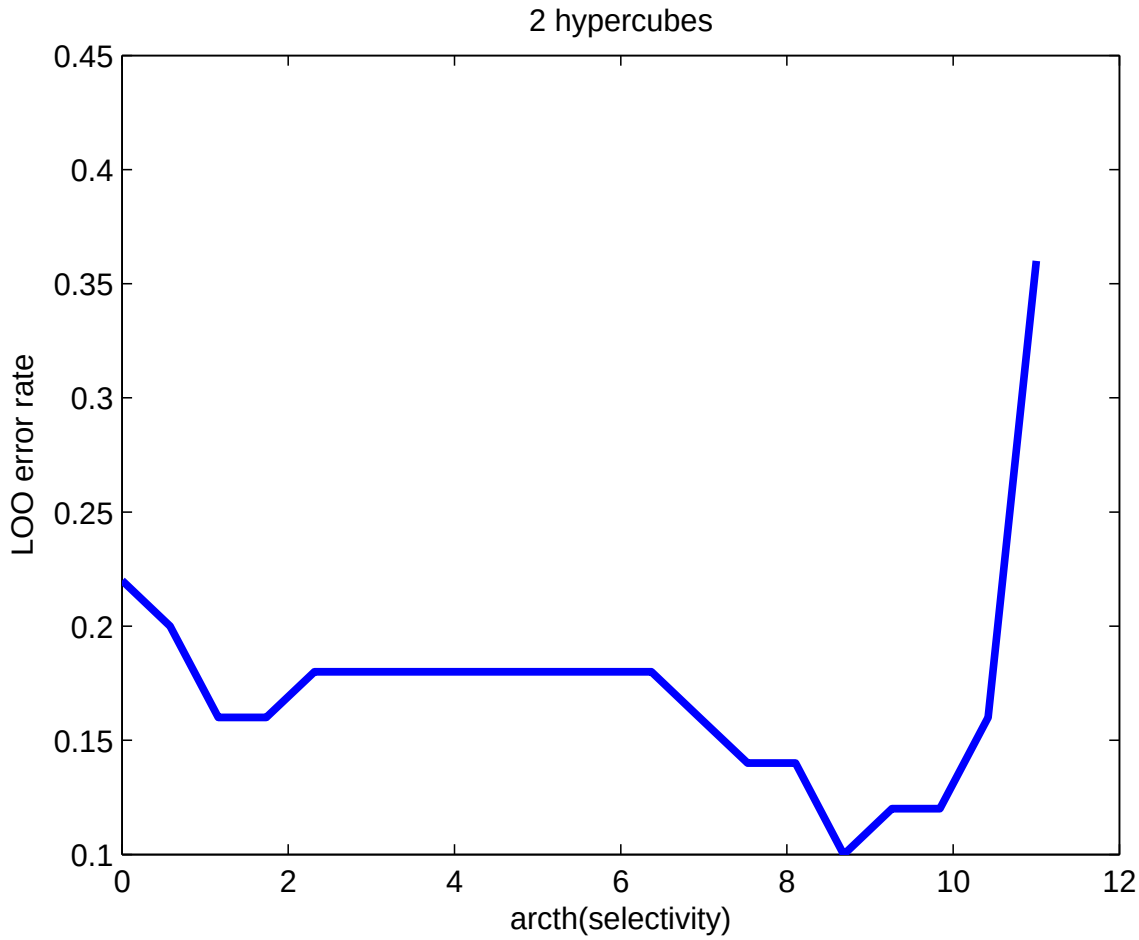

Рис. 3: Зависимость доли ошибок Leave One Out от значения параметра селективности

- В точке минимума ошибки алгоритм выделяет 13 признаков, содержащих 5 правильных.
- Отбрасываются 37 шумовых признаков.
- Использование метода опорных признаков позволяет уменьшить долю ошибок на Leave One Out на 12 % по сравнению с обычным SVM.

3.3 Реальные данные: Lung Cancer

- $n = 55$ - число признаков, $N_1 = 9$ - число больных пациентов, $N_2 = 18$ - число здоровых пациентов.

- Источник данных: интернет-репозиторий задач машинного обучения UCI.

3.4 Результаты на реальных данных

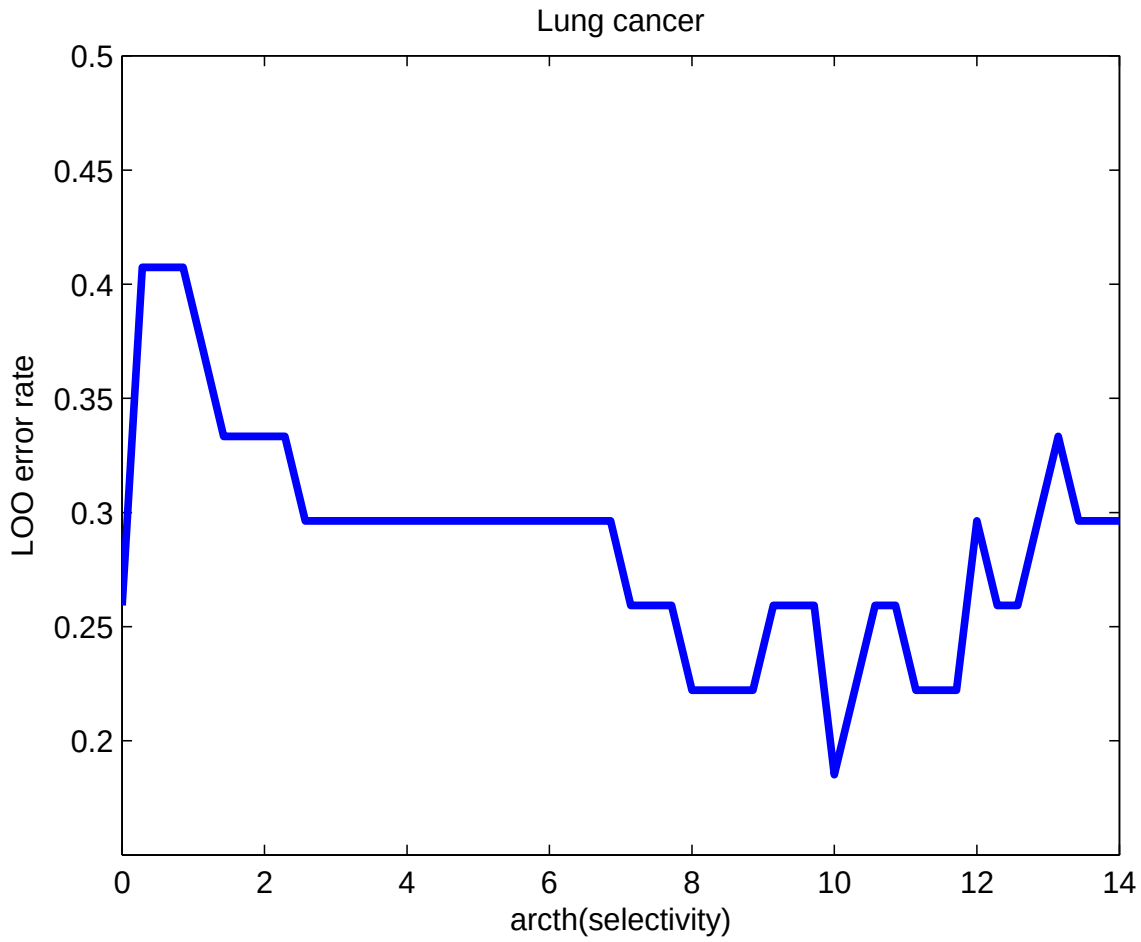

Рис. 4: Зависимость доли ошибок Leave One Out от значения параметра селективности

- В точке минимума ошибки алгоритм выделяет 5 признаков.
- Использование метода опорных признаков позволяет уменьшить долю ошибок на Leave One Out на 6 % по сравнению с обычным SVM.

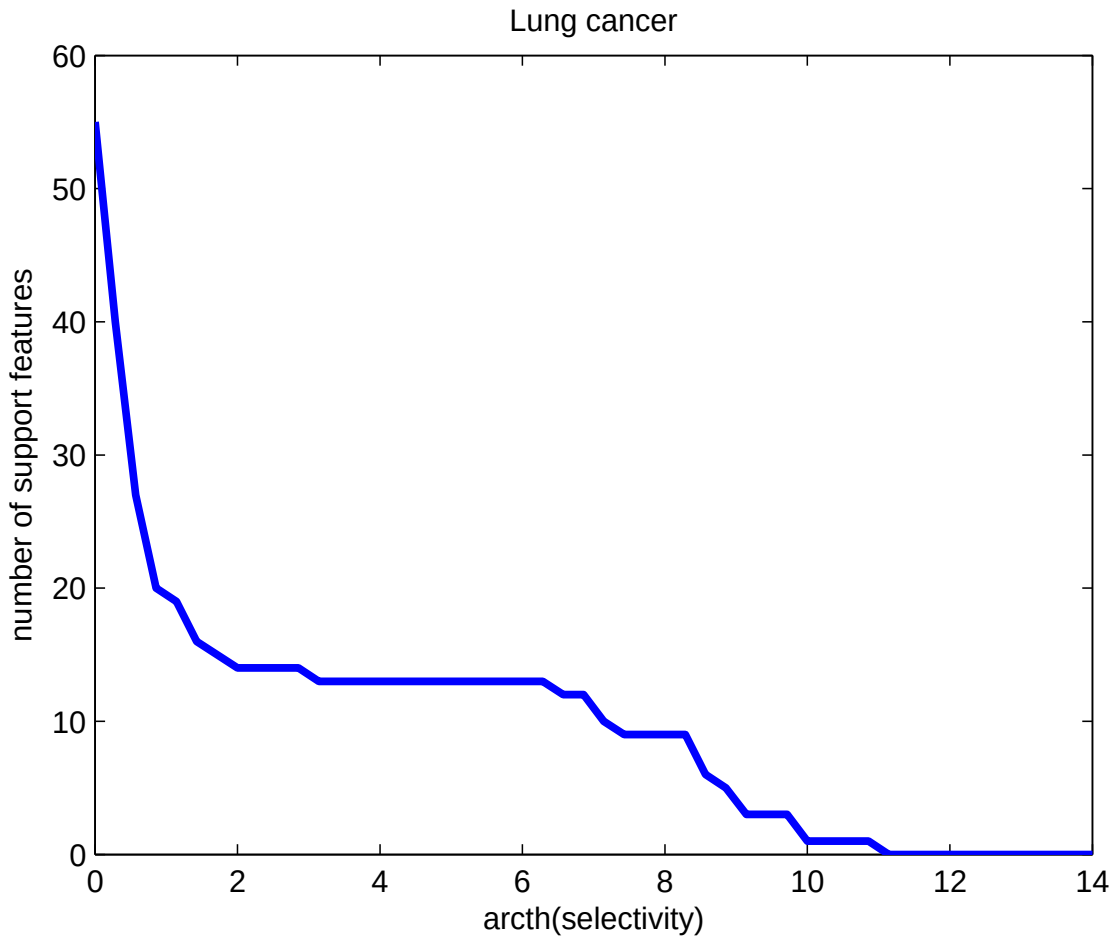

Рис. 5: Зависимость числа признаков от значения параметра селективности

- При приближении селективности к 1 алгоритм отбрасывает все признаки.

3.5 Обсуждение и выводы

Эксперименты подтверждают, что использование регуляризатора l_2-l_1 позволяет осуществлять отбор признаков в методе опорных векторов. Также экспериментально установлено, что параметр селективности μ регулирует количество признаков в модели. При $\mu = 0$ отбор не осуществляется и метод сводится к обычному SVM. При неограниченном приближении селективности к 1 отбрасываются все признаки. Оптимальное значение μ необходимо выбирать по внешнему критерию (Leave One Out в данной работе). Такой выбор позволяет существенно улучшить результаты метода

SVM за счет удаления из модели шумовых признаков, что подтверждено экспериментально: доля ошибок на Leave One Out уменьшилась на 6 % в эксперименте с реальными данными.

Изложенный метод опорных признаков имеет смысл применять к моделям с избыточным признаковым описанием, а именно, когда $n \geq N$. В этих случаях SVM склонен к переобучению, и уменьшение размерности задачи позволяет перейти к тем границам, в которых SVM работает эффективно.

При использовании метода опорных признаков возникает проблема эффективного численного решения двойственной задачи оптимизации. Если настройку параметров классического SVM можно осуществить за $O(N^2)$ действий по методу SMO, то в данном методе такой вычислительной эффективности достичь не удается. Сведение двойственной задачи к квадратичной позволяет решить задачу лишь за $O((N+n)^3)$ действий, что делает метод практически неприменимым в задачах классификации с числом признаков порядка сотен тысяч. Для решения данной проблемы предложена идея обобщения метода SMO на неквадратичный случай, что в перспективе позволит решать задачи указанных размеров.

4 Заключение

В данной работе получены следующие результаты:

- Предложен регуляризатор l_2-l_1 с параметром селективности для задачи обучения распознаванию объектов двух классов.
- Получена двойственная задача и выражение для решения исходной задачи через решение двойственной.
- Предложено два метода решения двойственной задачи: сведением к квадратичной задаче и с помощью обобщения метода SMO.
- Исследовано поведение метода на реальных и модельных данных с выбором селективности по Cross Validation.

Список литературы

[1] Vapnik V. N. *Statistical Learning Theory.* Wiley-Interscience, 1998.

[2] Platt J. C. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization.* MIT Press, 1999.

[3] Marron S. *An Overview of Support Vector Machines and Kernel Methods.* Talk, 2003.

[4] Tibshirani R. *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society, 1996.

[5] Сухарев А. Г., Тимохов А. В., Федоров В. В. *Курс методов оптимизации.* Физматлит, 2005.

[6] Nocedal J., Wright S. *Numerical Optimization.* Springer, 2006.