ВЕРОЯТНОСТНЫЕ МОДЕЛИ

УДК 681.513.7

О ДОСТИЖЕНИИ КОМПРОМИССА МЕЖДУ ТОЧНОСТЬЮ И УСТОЙЧИВОСТЬЮ КЛАССИФИКАТОРОВ В ЗАДАЧЕ ВЫБОРА НАИЛУЧШЕЙ ЯДРОВОЙ ФУНКЦИИ ПРИ БАЙЕСОВСКОМ ОБУЧЕНИИ[1]

Ветров Д.П.[*], **Кропотов Д.А.**[**], **Пташко Н.О.**[*]
[*] ВМиК МГУ им. М.В. Ломоносова, г. Москва
[**] ВЦ РАН, г. Москва

Поступила в редакцию 25.05.2009, после переработки 26.06.2009.

В данной работе показано, что подбор ядровой функции в методе релевантных векторов (RVM) требует расширения семейства классификаторов. В новой модели интегрирование по апостериорной плотности распределения становится вычислительно трудоемким, поэтому используется её точечное оценивание. Предложен метод локальной аппроксимации обоснованности, которая позволяет установить компромисс между точностью классификатора и его устойчивостью.

In the paper we show that RBF kernel selection in relevance vector machines (RVM) classifier requires extension of classifiers model. In new model integration over posterior probability becomes computationally unavailable. We propose a method of local evidence estimation which establishes a compromise between accuracy and stability of classifier.

Ключевые слова: распознавание образов, байесовский подход, выбор модели, метод релевантных векторов.
Keywords: machine learning, bayesian framework, model selection, relevance vector machine.

1. Введение

Рассматривается следующая задача классификации с двумя классами. Дано множество из m входных объектов $\{\vec{x}_i\}_{i=1}^m$ и соответствующие им метки классов $\{t_i\}_{i=1}^m$. Для удобства t_i принимает значения из множества $\{-1, 1\}$. Требуется, используя данную обучающую информацию, построить алгоритм для точного определения классов новых объектов. Одним из широко известных методов решения такой задачи классификации является метод опорных векторов (Support Vector Machine, SVM), производящий распознавание на основе следующей модели:

$$y(x) = \text{sign}(\sum_{i=1}^{m} w_i K(x, x_i) + w_0) \qquad (1)$$

[1]Работа выполнена при финансовой поддержке РФФИ (коды проектов 07-01-00211-а, 06-01-00492, 06-06-80464).

где $\{w_i\}_{i=0}^m$ - множество действительных переменных, «весов» прецедентов, а $K(\cdot,\cdot)$ - двухместная ядровая функция, обладающая свойствами симметричности и неотрицательной определенности.

SVM показал высокое качество работы на многих задачах [1]. Главные причины его успеха состоят в следующем. Идея Вапника о построении оптимальной разделяющей гиперплоскости привела к принципу максимизации «зазора» (*margin*) [2], который обеспечивает лучшую обобщающую способность. Вместе с этим, использование ядровых функций («kernel-trick» позволяет) применять линейные методы машинного обучения для построения нелинейных поверхностей. Успешное применение SVM требует адекватного выбора ядровой функции, а также константы регуляризации C, ограничивающей абсолютные значения весов при решении оптимизационной задачи. Различные значения C и формы ядровой функции приводят к существенно различным поведениям SVM в конкретных задачах.

Обычно параметры ядровых функций и коэффициент C определяются с использованием процедуры кросс-валидации (скользящий контроль), которая может быть слишком трудоемка с вычислительной точки зрения. Более того, оценка качества с помощью кросс-валидации, хотя и является несмещенной [2], может иметь большую дисперсию из-за ограниченности обучающей выборки.

Задача подбора ядровой функции вызывает острый интерес многих исследователей. Вапник и его единомышленники предполагают проводить подбор ядер теоретически, получая выражения для верхней границы ошибки на скользящем контроле. В частности, группой Вапника получены так называемый критерий R^2W^2 [3] и оценка, основанная на понятии «разброса опорных векторов» (*span of support vectors*) [4]. В [5] используется эволюционная стратегия настройки параметров с помощью процедуры дискретного поиска, которая требует больших временных затрат.

Одним из путей решения исследуемой задачи является использование Байесовского похода к обучению и предложенного Маккаем принципа максимума обоснованности (evidence) [6]. В этом случае, как правило, предлагается некоторая вероятностная интерпретация алгоритма распознавания.

М. Типпингом был предложен «метод релевантных векторов» (Relevance Vector Machine, RVM), также основанный на модели (1) и использующий Байесовскую регуляризацию для подбора весов алгоритма [7]. В данном алгоритме веса так называемых «релевантных» векторов интерпретируются как нормально распределенные случайные величины с нулевыми математическими ожиданиями. Предлагаемый подход не требует подбора коэффициента регуляризации C, ограничивающего значения весов, т.к. большие веса автоматически штрафуются во время обучения. Тем не менее, проблема подбора ядровой функции остается открытой. В данной работе предлагается метод подбора ядровой функции для RVM. Рассматривается наиболее популярное семейство ядровых функций - гауссовские потенциальные функции (RBF, Radial Basis Functions) и задача подбора параметра σ, отвечающего за ширину гауссианы. Ниже показывается, что применение Байесовского подхода к поставленной задаче требует расширения модели классификаторов, а именно, включения центров ядровых функций в параметры алгоритмов. При таком определении модели интегрирование по всему апостериорному распределению становится вычислительно сложным, а используемая в подобных ситуациях аппроксимация Лапласа (аппроксимация функции гауссианой в точ-

ке ее максимума) неадекватна в силу многоэкстремальности апостериорного распределения. Более того, нахождение ее максимума также является чрезвычайно трудоемкой процедурой из-за высокой размерности пространства оптимизации. В работе предлагается метод оценки локальной обоснованности, учитывающий указанные особенности задачи и приводящий к компромиссу между устойчивостью и точностью классификатора.

В главах 2 и 3 части 1 кратко описаны идеи Байесовского обучения, принципа максимума обоснованности и классификации методом релевантных векторов. В главе 4 части 1 изложен предлагаемый подход, а также показано, почему принцип максимума обоснованности не может быть напрямую использован для подбора ядровой функции. В главе 1 части 2 описан алгоритм подбора ядровой функции разработанный на основе введенного принципа устойчивости. Глава 2 части 2 содержит результаты экспериментов по применению предложенного подхода для модельных и реальных задач. В последней главе части 2 приведены заключение и выводы.

2. Байесовское обучение и принцип максимума обоснованности

Определение 1. Вероятностной моделью алгоритмов распознавания назовем тройку $\langle W, P(D|\vec{w}), P(\vec{w})\rangle$, где $W = \{\vec{w}\}$ - множество алгоритмов распознавания, $P(D|\vec{w})$ - функция правдоподобия выборки D при фиксированном алгоритме распознавания $\vec{w}$ и $P(\vec{w})$ - априорное распределение на W.

Предположим, что имеется параметрическое семейство вероятностных моделей $\{\langle W(\vec{\alpha}), P(D|\vec{w},\vec{\alpha}), P(\vec{w}|\vec{\alpha}), \vec{\alpha} \in A\rangle\}$, с заданным на нем априорным распределением $P(\vec{\alpha})$.

Представим правдоподобие тестовой выборки в виде интеграла:

$$P(D_{test}|D_{train}) = \int\limits_{A}\int\limits_{W(\vec{\alpha})} P(D_{test}|\vec{w},\vec{\alpha})P(\vec{w},\vec{\alpha}|D_{train})d\vec{w}d\vec{\alpha} = \tag{2}$$

$$= \int\limits_{A}\int\limits_{W(\vec{\alpha})} P(D_{test}|\vec{w})P(\vec{w}|\vec{\alpha},D_{train})P(\vec{\alpha}|D_{train})d\vec{w}d\vec{\alpha},$$

где

$$P(\vec{\alpha}|D_{train}) = \frac{1}{Z}P(D_{train}|\vec{\alpha})P(\vec{\alpha}), Z - \text{выражение, независящее от } \vec{\alpha}.$$

Интегрирование по A обычно затруднительно, поэтому $P(\vec{\alpha}|D_{train})$ часто аппроксимируется $\delta(\alpha_{\vec{MP}})$, где $\alpha_{\vec{MP}} = \arg\max\limits_{\vec{\alpha}} P(\vec{\alpha}|D_{train})$. Тогда равенство (2) переходит в

$$P(D_{test}|D_{train}) \approx \int\limits_{W(\alpha_{\vec{MP}})} P(D_{test}|\vec{w})P(\vec{w}|\alpha_{\vec{MP}},D_{train})d\vec{w}. \tag{3}$$

Определение 2. Назовем обоснованностью [6] модели $\langle W(\vec{\alpha}), P(D|\vec{w},\vec{\alpha}), P(\vec{w}|\vec{\alpha})\rangle$ величину

$$P(D_{train}|\vec{\alpha}) = \int\limits_{W(\vec{\alpha})} P(D_{train}|\vec{w})P(\vec{w}|\vec{\alpha})d\vec{w} \tag{4}$$

Заметим, что в случае отсутствия априорных предположений на $\vec{\alpha}$, верно

$$\arg\max_{\vec{\alpha}} P(\vec{\alpha}|D_{train}) = \arg\max_{\vec{\alpha}} P(D_{train}|\vec{\alpha})P(\vec{\alpha}) = \arg\max_{\vec{\alpha}} P(D_{train}|\vec{\alpha}).$$

Итак, процедура обучения сводится к поиску такой модели (то есть такого вектора $\vec{\alpha}$), для которой значение величины $P(D_{train}|\vec{\alpha})$, то есть обоснованности, максимально.

3. Метод релевантных векторов

Рассмотрим идею применения Байесовского подхода к методам потенциальных функций, предложенную Типпингом [7]. Рассматривается задача классификации с двумя классами. Пусть $D_{train} = \{\vec{x},\vec{t}\} = \{x_i,t_i\}_{i=1}^m$ - обучающая выборка, где $x_i = (x_i^1,\ldots,x_i^n)$ векторы из n-мерного вещественного признакового пространства, а t_i - метки классов, принимающие значения из множества $\{-1,1\}$. Рассмотрим семейство классификаторов

$$W = \left\{ y(x_{new}) = \mathrm{sign}(\sum_{i=1}^{m} w_i K(x_{new},x_i) + w_0) = \mathrm{sign}(h(x_{new},\vec{w})) \right\}.$$

Зададим параметрическое семейство моделей классификаторов

$$\left\{\langle W(\vec{\alpha}), P(D|\vec{w},\vec{\alpha}), P(\vec{w}|\vec{\alpha}), \vec{\alpha}\in A\rangle\right\}.$$

Положим $W(\vec{\alpha}) = W, A = \mathbb{R}^{m+1}$. Установим априорную вероятность на веса

$$P(w_i|\alpha_i) \sim N(0,\alpha_i^{-1}), i = 1,...,m+1.$$

Определим правдоподобие обучающей выборки как

$$P(D_{train}|\vec{w},\vec{\alpha}) = P(D_{train}|\vec{w}) = \prod_{i=1}^{m} \frac{1}{1+\exp(-t_i h(x_i,\vec{w}))}$$

Обоснованность модели задается выражением (4). Наша цель состоит в нахождении $\vec{\alpha}$, максимизирующего обоснованность, и в последующем получении апостериорного распределения $P(\vec{w}|\vec{\alpha},D_{train})$. Обозначим для краткости $Q_{\vec{\alpha}}(\vec{w}) = P(D_{train}|\vec{w})P(\vec{w}|\vec{\alpha})$. Т.к. прямое вычисление (4) затруднительно, Типпинг использовал аппроксимацию Лапласа, раскладывая $L_{\vec{\alpha}}(\vec{w}) = \log Q_{\vec{\alpha}}(\vec{w})$ по $\vec{w}$ в ряд Тейлора в точке максимума $w_{\vec{MP}}$ и аппроксимируя ее квадратичной функцией. Полученная функция может быть проинтегрирована аналитически.

$$P(D_{train}|\vec{\alpha}) \approx Q_{\vec{\alpha}}(w_{\vec{MP}}) \mid \Sigma \mid^{1/2}, \tag{5}$$

$$\Sigma = (-\nabla_{\vec{w}}\nabla_{\vec{w}}L_{\vec{\alpha}}(\vec{w})\mid_{\vec{w}=w_{\vec{MP}}})^{-1} = (-\nabla_{\vec{w}}\nabla_{\vec{w}}\log(P(D_{train}\mid\vec{w})) - A)^{-1} \quad (6)$$

где $A = diag(\alpha_1, \ldots, \alpha_m)$. Дифференцируя последнее выражение по $\vec{\alpha}$ и приравнивая производные к нулю получаем формулу для пересчета $\vec{\alpha}$

$$\alpha_i^{new} = \frac{\gamma_i}{w_{MP,i}^2} \quad (7)$$

$$\gamma_i = 1 - \alpha_i^{old}\Sigma_{ii} \quad (8)$$

Здесь γ_i - это так называемый эффективный вес i-ого параметра, показывающий, насколько его значение ограничено коэффициентом регуляризации α_i, определяющей априорное распределение параметра. Легко показать, что $\gamma_i \in [0, 1]$. Если α_i близко к нулю, то априорное распределение практически не влияет на значение w_i, и γ_i стремится к единице. Напротив, в случае больших α_i соответствующий параметр близок к нулю и не зависит от обучающей информации, т.е. его эффективный вес стремится к нулю.

Процедура обучения RVM состоит из трех итеративных шагов. Сначала ищется точка максимума $w_{\vec{MP}}$ функции $L(\vec{w})$. Затем по схеме (5) проводится аппроксимация и, используя (7), находятся новые значения параметров $\vec{\alpha}$. Шаги повторяются до тех пор, пока процесс не сойдется.

После окончания обучения, интеграл (3) может быть заменен точечной оценкой. Полагая $P(\vec{w}|D_{train}, \vec{\alpha}) \approx \delta(w_{\vec{MP}})$ получаем выражение

$$P(D_{test}|D_{train}) = P(D_{test}|w_{\vec{MP}}) \quad (9)$$

В [7] показано, что RVM обладает примерно таким же качеством, как и SVM со значением параметра C, выбранным с помощью процедуры скользящего контроля, при этом не требуя ручной установки параметров регуляризации. Более того, оказалось, что RVM является намного более разреженным, т.е. число ненулевых весов (релевантных векторов) значительно меньше количества опорных векторов. Это происходит вследствие того, что большинство объектов оказываются нерелевантными, и соответствующие α стремятся к бесконечности.

4. Подбор ядровой функции для RVM

Хотя принцип максимальной обоснованности сформулирован полностью в вероятностных терминах, для него можно предложить иную интерпретацию. Выражение (5) может быть рассмотрено как компромисс между точностью алгоритма на обучающей выборке (значение $Q_{\vec{\alpha}}(w_{\vec{MP}})$) и устойчивостью относительно малых изменений параметров (выраженной корнем обратной величины определителя гессиана $\log(Q_{\vec{\alpha}}(w_{\vec{MP}}))$). Используя данные понятия, можно сформулировать *принцип устойчивости*: обобщающая способность классификатора тем выше, чем он более "устойчив". В данном случае термин устойчивость понимается довольно неформально. Различные его определения рассматривались многими исследователями [8], [9]. Применительно к Байесовскому обучению устойчивость понимается как способность сохранять высокое правдоподобие (или более точно $Q_{\vec{\alpha}}(\vec{w})$) при изменении параметров алгоритма в точке максимума. Такой взгляд позволяет изменить концепцию Байесовской регуляризации в тех случаях, когда прямое ее применение невозможно или бессмысленно.

Выбор модели с помощью принципа максимума обоснованности позволяет избежать прямой установки ограничений на веса в RVM. Тем не менее, вопрос о выборе подходящей ядровой функции остается открытым. Далее будем рассматривать популярное семейство ядровых функций $K(x,z) = \exp(-\frac{||x-z||^2}{2\sigma^2})$. Наша цель состоит в нахождении наилучшего значения параметра σ без трудоемкой процедуры скользящего контроля, используя принцип устойчивости.

Будем интерпретировать тип ядровой функции как параметр модели и использовать принцип устойчивости для его выбора. Учитывая, что параметры σ связаны с устойчивостью относительно сдвига центров ядровых функций, для подбора значений σ необходимо расширить модель (1), включив в нее, в качестве параметров, центры ядровых функций: $h_E(x_{new}, \vec{w}, \vec{z}) = \text{sign}(\sum_{i=1}^{m} w_i K(x_{new}, z_i) + w_0)$. В новой модели прямое вычисление обоснованности (4) становится невозможным. Для использования аппроксимации Лапласа требуется оптимизация положения центров ядровых функций z_i

$$L_{\sigma,\vec{\alpha}}(\vec{w}, \vec{z}) = \log(P(D_{train}|\vec{w}, \vec{z})P(\vec{w}|\vec{\alpha})P(\vec{z})) \to \max_{\vec{w},\vec{z}} \tag{10}$$

К сожалению, оптимизация $L_{\sigma,\vec{\alpha}}(\vec{w}, \vec{z})$ по $\vec{z}$ затруднительна вследствие высокой размерности пространства оптимизации, т.к. $\vec{z} \in \mathbb{R}^{mn}$. Кроме того, $L_{\sigma,\vec{\alpha}}(\vec{w}, \vec{z})$ многоэкстремально по $\vec{z}$. По этой причине вместо разложения $L_{\sigma,\vec{\alpha}}(\vec{w}, \vec{z})$ в ряд Тейлора в точке максимума $(\vec{w}_{MP}, \vec{z}_{MP})$ (для аппроксимации Лапласа) будем использовать разложение в точке $(\vec{w}_{MP}, \vec{z})$, где $\vec{w}_{MP}$ - точка максимума $L_{\sigma,\vec{\alpha}}(\vec{w}, \vec{z})$ при $\vec{z}$ фиксированных в объектах обучающей выборки. Заметим, что в этом случае производная $\frac{\partial L_{\sigma,\vec{\alpha}}(\vec{w},\vec{z})}{\partial z_i} \mid_{z_i=x_i} \neq 0$, и, таким образом, первая и вторая производные становятся важны для определения меры устойчивости относительно гипотетических перемещений центров ядровых функций. Введем локальный аналог аналог обоснованности, на основе которого будем проводить выбор модели. Для этого обобщим выражение (5) на случай многоэкстремального функционала (10) и точки, не являющейся экстремумом.

Будем оценивать устойчивость уже обученного классификатора относительно сдвига центров ядровых функций. Следовательно, можно считать веса $w_{\vec{MP}}$ константами, не зависящими от $\vec{z}$. В качестве априорного распределения на расположение центров возьмем несобственное равномерное распределение $P(\vec{z}) = const$. Далее перепишем выражение (10) в виде:

$$L_{\sigma,\vec{\alpha}}(\vec{w}, \vec{z}) = \log(P(D_{train}|\vec{w}, \vec{z})) + \Psi(\vec{w}, \vec{\alpha}) \tag{11}$$

где $\Psi(\vec{w}, \vec{\alpha})$ не зависит от $\vec{z}$ и, следовательно, может быть опущено при дифференцировании.

Обозначим через A_i устойчивость $P(D_{train}|w_{\vec{MP}}, \vec{z})$ по отношению к положению центра z_i. Далее, предположим, что она может быть разложена так, как если бы устойчивости по разным координатам были независимы

$$A_i = \prod_{j=1}^{n} A_{ij}$$

где A_{ij} выражает устойчивость классификатора относительно малых изменений j-ой координаты i-ого центра. Для определения A_{ij} аппроксимируем $\log(P(D_{train}|$

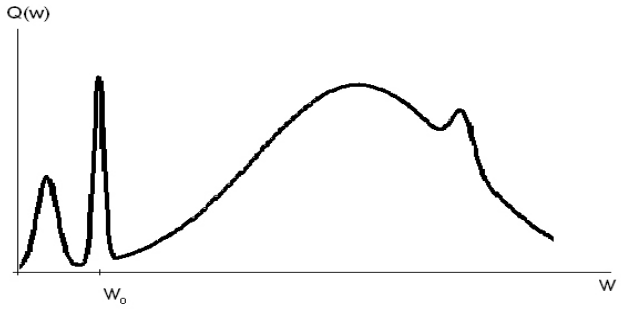

Рис. 1: Пример модели с высокой обоснованностью, но низкой обобщающей способностью наилучшего (в смысле регуляризованного правдоподобия) алгоритма. Высокое значение обоснованности модели не имеет смысла при использовании единственного алгоритма $(\vec{w},\vec{z}) = (w_{\vec{MP}}, z_{\vec{MP}})$. В то же время, локальные характеристики точки $(w_{\vec{MP}}, z_{\vec{MP}})$, такие как $\nabla_{w,z}\nabla_{w,z}Q(\vec{w},\vec{z})\mid_{(\vec{w},\vec{z})=(w_{\vec{MP}},z_{\vec{MP}})}$ понижают локальную обоснованность данного алгоритма.

$w_{\vec{MP}},\vec{z})$), разложив функцию в ряд Тейлора в точке $\vec{z}=\vec{x}$ по z_i^j. Получим

$$A_{ij} = \begin{cases} |a|^{-1}, & \text{если } b \le 0 \\ \frac{1}{2}\sqrt{\frac{2\pi}{b}}\exp\left(\frac{a^2}{2b}\right)\left(1-\operatorname{erf}\left(\frac{|a|}{\sqrt{2b}}\right)\right), & \text{иначе} \end{cases} \tag{12}$$

здесь

$$a = \frac{\partial \log(P(D_{train}|w_{\vec{MP}},\vec{z}))}{\partial z_i^j}$$

$$b = -\frac{\partial^2 \log(P(D_{train}|w_{\vec{MP}},\vec{z}))}{(\partial z_i^j)^2}$$

Смысл выражения (12) показан на фигуре . Сначала аппроксимируем логарифм функции $f(z_i^j) = P(D_{train}|w_{\vec{MP}},\vec{z})$ параболой (с отрицательным коэффициентом при квадрате), или прямой (если вторая производная не отрицательна). Аппроксимацию проводим в точке $z_i^j = x_i^j$. Получается приближение убывающей части $Q(\vec{w},\vec{z})$ гауссианой или экспонентой. Далее проводим интегрирование для того, чтобы получить меру устойчивости в терминах производных. Отметим, что эти действия проводятся по аналогии с (5). Если x_i^j была бы точкой экстремума $f(z_i^j)$, тогда A_{ij} с точностью до константного множителя совпало бы со значением приближения Лапласа функции $Q(\vec{w},\vec{z})$ по координате x_i^j.

Далее можно объединить устойчивость и точность в одном выражении. Для этого рассмотрим вес каждого центра ядровой функции. Если вес центра близок к нулю, то его устойчивость не должна играть роли. Ясно, что нулевые веса w_i соответствуют тому случаю, когда в z_i отсутствует центр. Рассматривая эффективные веса (8) каждого центра γ_i, изменяющиеся от 0 до 1, получаем выражение

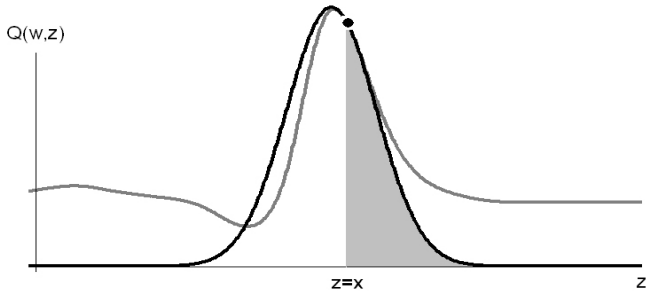

Рис. 2: Оптимизация в пространстве центров $\vec{z}$ затруднительна, поэтому центры ядровых функций помещены в объектах обучающей выборки, т.е. $\vec{z} = \vec{x}$. Это может привести к ненулевому значению градиента $\nabla_{\vec{z}} Q(\vec{w}, \vec{z})$. Для того, чтобы установить компромисс между точностью данного алгоритма и его способностью сохранять свое качество при изменении параметров, достаточно проинтегрировать убывающую часть лапласовского приближения функции $Q(\vec{w}, \vec{z})$ (серая область). Получаем выражение, сочетающее локальную точность алгоритма (значение $Q(\vec{w}, \vec{z})$) и его устойчивость (значения производных).

для полной устойчивости правдоподобия относительно всех центров

$$Z = \prod_{i=1}^{m} A_i^{\gamma_i} = \prod_{i=1}^{m} (\prod_{j=1}^{n} A_{ij})^{\gamma_i} \tag{13}$$

Последнее условие эквивалентно взвешенной сумме при взятии логарифма

Теорема 1. Выражение (13) определено корректно в смысле добавления или удаления ядер с нулевым весом.

Пусть $w_1 = 0$. Тогда докажем, что

$$\prod_{i=1}^{m} A_i^{\gamma_i} = \prod_{i=2}^{m} A_i^{\gamma_i}$$

При обучении RVM максимизируется обоснованность (5). Если $w_1 = 0$, то можно показать, что значение α_1 стремится к бесконечности. В противном случае, вес w_1 хотя бы немного настроился на данные. Эффективный вес выражается следующей формулой: $\gamma_1 = 1 - \alpha_1 \Sigma_{11}$, где Σ_{11} - первый элемент обратного Гессиана (6)

$$\Sigma = (-\nabla_{\vec{w}} \nabla_{\vec{w}} L_{\vec{\alpha}}(\vec{w}) \mid_{\vec{w}=w_{\vec{MP}}})^{-1}$$

Для больших α_1 справедливо, что $\Sigma_{11} \approx \alpha_1^{-1}$. Таким образом, предел

$$\lim_{\alpha_1 \to \infty} \alpha_1 \Sigma_{11} = 1$$

и, следовательно, $\gamma_1 = 0$. Подставляя полученный результат в (13) имеем

$$\prod_{i=1}^{m} A_i^{\gamma_i} = A_1^{\gamma_1} \prod_{i=2}^{m} A_i^{\gamma_i} = \prod_{i=2}^{m} A_i^{\gamma_i}$$

Это означает, что можно добавлять или удалять произвольное число центров с нулевым весом, не меняя при этом выражение для устойчивости. Теорема доказана.

Определение 3. Перемножая устойчивость алгоритма Z и значение правдоподобия в точке $w_{\vec{MP}}$, получаем значение для *коэффициента ядровой пригодности*

$$KV = P(D_{train}|w_{\vec{MP}}, \vec{z})Z \tag{14}$$

Таким образом, при интерпретации обоснованности как соотношения между точностью и устойчивостью классификатора (4), можно получить аналогичное выражение для более общего случая. Итак, в качестве итоговой ядровой функции для данной задачи обучения предлагается взять функцию, которой соответствует наибольшее значение коэффициента ядровой пригодности.

Список литературы

[1] Burges, C.J.S: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2 (1998) 121–167

[2] Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag New York (1995)

[3] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., Vapnik, V.: Feature Selection for Support Vector Machines. Proc. of 15th International Conference on Pattern Recognition, Vol.2 (2000)

[4] Chapelle, O., Vapnik, V.: Model Selection for Support Vector Machines. Advances in Neural Information Processing Systems 12, ed. S.A. Solla, T.K. Leen and K.-R. Muller, MIT Press (2000)

[5] Ayat, N.E., Cheriet, M., Suen, C.Y.: Optimization of SVM Kernels using an Empirical Error Minimization Scheme. Proc. of the First International Workshop on Pattern Recognition with Support Vector Machines (2002)

[6] MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press (2003)

[7] Tipping, M.E.: Sparse Bayesian Learning and the Relevance Vector Machines. Journal of Machine Learning Research 1 (2001) 211–244

[8] Kutin, S., Niyogi, P.: Almost-everywhere algorithmic stability and generalization error. Tech. Rep. TR-2002-03: University of Chicago (2002)

[9] Bousquet, O., Elisseeff, A.: Algorithmic stability and generalization p erformance. Advances in Neural Information Processing Systems 13 (2001)