На правах рукописи

Саутин Александр Сергеевич

РАЗРАБОТКА И ИССЛЕДОВАНИЕ МЕТОДОВ ПОСТРОЕНИЯ РЕГРЕССИОННЫХ МОДЕЛЕЙ НА ОСНОВЕ АЛГОРИТМА ОПОРНЫХ ВЕКТОРОВ И ЕГО МОДИФИКАЦИЙ

05.13.17 – Теоретические основы информатики

Автореферат диссертации на соискание ученой степени кандидата технических наук

Новосибирск – 2010

Работа выполнена в Государственном образовательном учреждении высшего профессионального образования «Новосибирский государственный технический университет»

Научный руководитель:	доктор технических наук, профессор Попов Александр Александрович
Официальные оппоненты:	доктор технических наук, профессор Загоруйко Николай Григорьевич
	кандидат технических наук, доцент Фаддеенков Андрей Владимирович
Ведущая организация:	Государственное образовательное учреждение высшего профессионального образования «Томский государственный университет систем управления и радиоэлектроники», г. Томск

Защита состоится « 17 » декабря 2010 г. в 10^{00} часов на заседании диссертационного совета Д 212.173.06 при Государственном образовательном учреждении высшего профессионального образования «Новосибирский государственный технический университет» (630092, Новосибирск-92, пр. К. Маркса, 20).

С диссертацией можно ознакомиться в библиотеке Новосибирского государственного технического университета.

Автореферат разослан « 16 » ноября 2010 г.

Ученый секретарь
диссертационного совета Чубич В.М.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследований. Задача восстановления зависимостей по эмпирическим данным была и, вероятно, всегда будет одной из главных в прикладном анализе. Эта задача является математической интерпретацией одной из основных проблем естествознания: как найти существующую закономерность по разрозненным фактам.

В наиболее общей постановке проблема восстановления зависимости приводит к задаче подбора модели оптимальной сложности. Изначально данная задача была как бы внешней и не встраивалась сразу в одну общую задачу. Примером нового подхода является подход по самоорганизации моделей, в свое время развитый школой А. Г. Ивахненко, а впоследствии и А. А. Поповым, принесшим в него идеи оптимального планирования эксперимента, в частности, разбиения выборки на обучающую и проверочную, в целом – идею активной структурной идентификации. Пожалуй, одним из первых подходов, когда организуется одна общая задача, в параметрическом случае является метод LASSO, предложенный Р. Тибширани. В непараметрическом случае одним из подходов является алгоритм опорных векторов (Support Vector Machines – SVM).

Изначально SVM был использован для решения задачи классификации данных. Позже, в 1996 году В. Вапником, Х. Драккером, К. Берджесом, Л. Кауфман и А. Смолой была предложена модификация SVM применительно к задаче построения регрессионных моделей. Метод SVM активно развивался в последующие годы такими учеными как А. Смола, Дж. Сайкенс, К. Кортес, Т. Джоатимс и др.

За небольшой промежуток времени алгоритм опорных векторов был использован для решения задач классификации данных и восстановления зависимостей во многих областях. Особенно успешным его применение было в таких областях как распознавание лиц, категоризация текстов, построение регрессионных моделей, предсказание временных рядов и распознавание рукописных символов.

При восстановлении зависимостей изначально в SVM использовалась функция потерь Вапника, которая представляет собой расширение функции потерь Лапласа путем добавления зоны нечувствительности. Впоследствии Дж. Сайкенсом было предложено расширение SVM, где использовалась квадратичная функция потерь (Гаусса). Данная модификация SVM получила название LS-SVM. Подробное исследование LS-SVM в задаче построения регрессионных моделей было проведено Дж. Бранбантером. Исследования LS-SVM в условиях автокорреляции ошибок наблюдений проводились М. Эспинозой, Дж. Сайкенсом и Б. Де Муром. Подробные исследования аппарата ядерных функций, предложенного М. А. Айзерманом, который позволил расширить применение SVM для восстановления нелинейных зависимостей, проводились А. Смолой, Б. Шелкопфом и К. Берджесом. Также в этой области исследований активно работали Н. Кристианини, Дж. Шов-Тейлор и др.

В связи с тем, что SVM сравнительно недавно разработанный метод, остается целый ряд вопросов его применения в задаче построения регрессионных

моделей. Этот ряд вопросов включает в себя использование SVM при различных моделях ошибок наблюдений, в условиях мультиколлинеарности данных, при нарушении предположений о независимости и постоянстве дисперсии ошибок наблюдений.

Цель и задачи исследований. Основной целью диссертационной работы является дальнейшее развитие, на основе использования компьютерного моделирования, SVM в задачах построения регрессионных моделей, и разработка его модификаций для более адекватного описания реальной ситуации.

В соответствии с поставленной целью решались следующие задачи:

– исследование возможностей использования SVM при построении регрессионных моделей в условиях наличия сильных выбросов в данных;
– разработка модификаций SVM для учета асимметричности ошибок наблюдений;
– разработка методов построения разреженных решений на основе SVM;
– исследование SVM в условиях мультиколлинеарности данных;
– построение модификаций SVM, направленных на возможность учета гетероскедастичности и автокорреляции ошибок наблюдений;
– разработка на основе SVM методов для построения квантильной регрессии и оценок неизвестной дисперсии ошибок наблюдений;
– разработка эффективных методов выбора гиперпараметров SVM.

Методы исследований. Для решения поставленных задач использовался аппарат теории вероятностей, математической статистики, вычислительной математики, математического программирования, статистического моделирования.

Научная новизна диссертационной работы заключается в:

– формулировках двойственных задач SVM для применения данного метода в условиях наличия сильных выбросов в данных и асимметричного засорения;
– результатах исследования SVM при асимметричных распределениях ошибок наблюдений и обобщении модификации SVM для построения квантильной регрессии на случай произвольной функции потерь;
– модификациях SVM для: получения разреженных решений, учета эффекта гетероскедастичности и автокорреляции ошибок наблюдений;
– результатах численных исследований предложенных методов с использованием технологии статистического моделирования.

Основные положения, выносимые на защиту.

1. Формулировки двойственных задач SVM при использовании адаптивных функций потерь и алгоритмы их решения.
2. Результаты исследования SVM в условиях асимметричных распределений ошибок наблюдений.
3. Расширение возможностей SVM при построении разреженных решений за счет использования адаптивных функций потерь.
4. Результаты исследования возможности использования SVM в условиях мультиколлинеарности данных, гетероскедастичности и автокорреляции ошибок наблюдений, а также при построении параметрических и полупараметрических моделей.

5. Результаты исследования возможности использования квантильного варианта SVM для построения доверительных интервалов и оценки неизвестной дисперсии.

Обоснованность и достоверность научных положений, выводов и рекомендаций обеспечивается:

– корректным применением аналитических методов исследования свойств построенных моделей;
– подтверждением аналитических выводов и рекомендаций результатами статистического моделирования.

Личный творческий вклад автора заключается в проведении исследований, обосновывающих основные положения, выносимые на защиту.

Практическая ценность и реализация результатов. Разработанные модификации SVM позволяют строить регрессионные модели в условиях наличия выбросов в данных и асимметричных распределений ошибок наблюдений. Предложенные методы на основе адаптивных функций потерь позволяют получать разреженные модели при использовании SVM на выборках данных большого объема. Проведенные исследования позволяют корректно использовать SVM в условиях мультиколлинеарности данных, а также в условиях гетероскедастичности и автокорреляции ошибок наблюдений. Созданное программное обеспечение позволяет эффективно строить регрессионные модели, применяя разработанные подходы.

Апробация работы. Основные результаты исследований, проведенных автором, докладывались и обсуждались на Российской НТК «Информатика и проблемы телекоммуникаций» (Новосибирск, 2008 и 2010); Всероссийской конференции студентов, аспирантов и молодых ученых «Молодежь и современные информационные технологии» (Томск, 2008); Третьем международном форуме по стратегическим технологиям IFOST (Новосибирск, 2008); Четвертом международном форуме по стратегическим технологиям IFOST (Хошимин, 2009); IX международной конференции «Актуальные проблемы электронного приборостроения АПЭП-2008» (Новосибирск, 2008).

Публикации. Основные научные результаты диссертации опубликованы в 11 печатных работах, из которых 2 – в журналах, рекомендованных ВАК, одна – в докладах АН ВШ РФ, 5 – в сборниках научных работ, 3 – в материалах конференций.

Структура работы. Диссертация состоит из введения, пяти глав, заключения, списка использованных источников (106 наименований) и двух приложений. Общий объем диссертации составляет 177 страниц, включая 21 таблицу и 58 рисунков.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Глава 1. Постановка задач исследования

В п. 1.1 представлена формулировка задачи построения математической модели явления.

Пусть имеются результаты n наблюдений за некоторой измеряемой величиной y. Известно, что свои значения она принимает в зависимости от набора входных данных x, которые в каждом из n опытов известны. В этом случае регрессионная модель наблюдения может быть записана в виде

$$y_i = r(x_i) + x_i,$$

где $y_i \in Y \subseteq R$ – i-ое наблюдение; $x_i \in X \subseteq R^d$ – значение входных данных в i-ом эксперименте; $r(x)$ – неизвестная функция; x_i – случайная ошибка.

Задача состоит в том, чтобы, располагая значениями входных данных и результатами проведенных наблюдений за измеряемой величиной y, как можно точнее оценить зависимость $r(x)$. Оценка этой зависимости производится на основе конечного числа наблюдений $(x_1, y_1), \mathbf{K}, (x_n, y_n) \in X \times Y$, где n – общее число наблюдений.

Для решения поставленной задачи используется теория машинного обучения. В п. 1.2 приводится краткое описание данной теории.

В п. 1.3 описывается алгоритм опорных векторов, который используется для решения поставленной задачи.

В п. 1.4 рассматривается вопрос получения разреженных решений на основе алгоритма опорных векторов.

В п. 1.5 описывается метод ядерных функций, который используется для расширения SVM на нелинейный случай.

Задача, возникающая при использовании алгоритма опорных векторов, является задачей квадратичного программирования. В п. 1.6 приводится обзор подходов к решению данной оптимизационной задачи.

Алгоритм опорных векторов имеет ряд настраиваемых параметров. В п. 1.7 приводится обзор подходов к выбору этих параметров.

В п. 1.8 представлен исторический обзор алгоритма опорных векторов. Проанализированы основные существующие на данный момент недостатки алгоритма.

В п. 1.9 обосновываются задачи исследований.

Глава 2. Конструирование двойственной задачи SVM с адаптивными функциями потерь

В данной главе приводится формулировка двойственной задачи для предложенных адаптивных функций потерь.

В п. 2.1 описывается методика построения квазиоптимальных функций потерь.

В п. 2.2 предлагаются адаптивные функции потерь для построения регрессионных моделей в условиях асимметричных засорений. За основу предлагаемой адаптивной функции потерь берется функция потерь Хьюбера. При этом изменяется второй сегмент этой функции. Он определяется как линейная комбинация линейно-возрастающей функции и горизонтальной прямой. Такая модификация дает возможность посредством скалярного параметра t изменять угол наклона линейного участка функции потерь, что снижает ее чувствитель-

ность к асимметричных засорениям. Скомбинированная функция потерь имеет вид:

$$L(x)=\begin{cases}\dfrac{1}{2s}x^2, & |x|\le s,\\ t\left(x-\dfrac{1}{2}s\right)+(1-t)\dfrac{1}{2}s, & |x|>s,\end{cases} \tag{1}$$

где t – угол наклона линейного участка функции потерь, s – параметр функции потерь Хьюбера, x – невязка.

Другим вариантом параметризации является использование в качестве основы функции потерь Лапласа. В этом случае функция потерь принимает вид:

$$L(x)=\begin{cases}|x|, & |x|\le s,\\ t|x|+(1-t)s, & |x|>s.\end{cases} \tag{2}$$

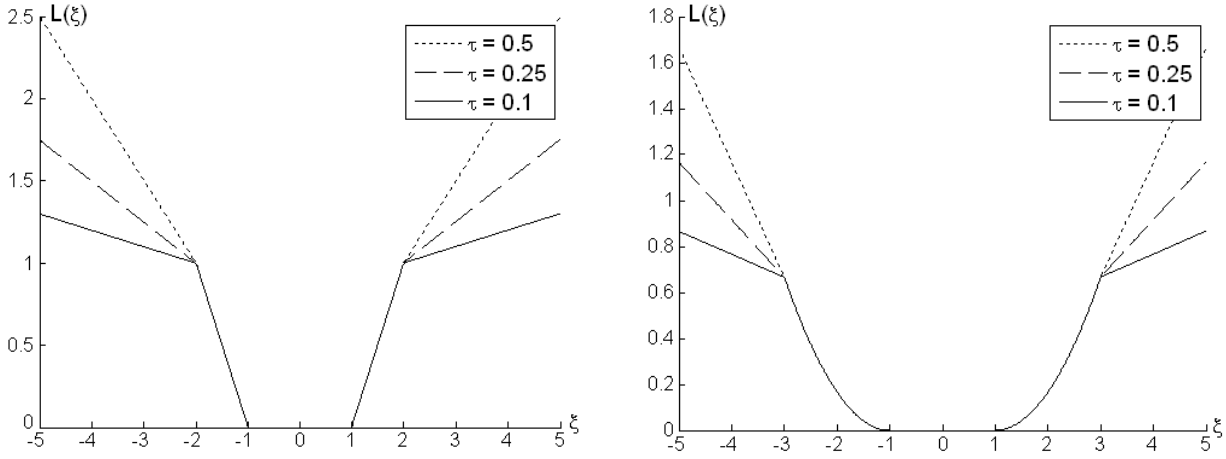
По аналогии с функцией потерь Вапника в предложенные адаптивные функции потерь можно добавить зону ε-нечувствительности. Тогда адаптивная функция Лапласа будет определяться выражением

$$L(x)=\begin{cases}0, & |x|\le e,\\ |x-e|, & e\le|x|\le s,\\ t|x-e|+(1-t)(s-e), & |x|>s,\end{cases} \tag{3}$$

а адаптивная функция Хьюбера – выражением

$$L(x)=\begin{cases}0, & |x|\le e,\\ \dfrac{1}{2s}(x-e)^2, & e\le|x|\le s,\\ t|x-e|+(1-t)(s-e), & |x|>s.\end{cases} \tag{4}$$

Здесь коэффициент e определяет ширину зону нечувствительности. Графики адаптивных функции потерь с зоной ε-нечувствительности показаны на рис. 1.



а) б)

Рис. 1. Вид адаптивных функций потерь при различных τ:
а) Лапласа (ε = 1, σ = 2); б) Хьюбера (ε = 1, σ = 3)

Стоит отметить, что адаптивные функции потерь Хьюбера и Лапласа в общем случае не являются выпуклыми функциями. Однако данные функции на интервале $t \in (0,1]$ удовлетворяют определению строго квазивыпуклой функции. Для задачи квадратичного программирования со строго квазивыпуклой целевой функцией и множеством ограничений, которое является выпуклым, локальный оптимум является глобальным и единственным. Следовательно, задача квадратичного программирования, возникающая при использовании данных функций потерь в алгоритме опорных векторов, также как и в случае с выпуклой функцией потерь, будет иметь единственный локальный минимум, и можно перейти к ее двойственной формулировке, и использовать для ее решения обычный метод оптимизации подобных задач.

В п. 2.3 формулируются двойственные задачи для предложенных адаптивных функций потерь. Для адаптивной функции потерь Лапласа задача принимает вид:

$$\max_{a_i, a_i^*} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_i - a_i^*\right)\left(a_j - a_j^*\right) j^T(x_i) j\left(x_j\right) + \sum_{i=1}^{n}\left(a_i - a_i^*\right) y_i$$

$$-e\sum_{i=1}^{n}\left(a_i + a_i^*\right) - C\sum_{i=1}^{n}\left(T\left(a_i\right) + T\left(a_i^*\right)\right), \qquad (5)$$

$$T\left(a_i\right) = \begin{cases} 0, & x_i \le s, \\ (1-t)s, & x_i > s, \end{cases} \qquad T\left(a_i^*\right) = \begin{cases} 0, & x_i^* \le s, \\ (1-t)s, & x_i^* > s \end{cases}$$

с ограничениями

$$a_i \in \begin{cases} [0, C], & x_i \le s, \\ [0, Ct], & x_i > s, \end{cases} \qquad a_i^* \in \begin{cases} [0, C], & x_i^* \le s, \\ [0, Ct], & x_i^* > s. \end{cases} \qquad (6)$$

Для адаптивной функции потерь Хьюбера меняются лишь выражения

$$T\left(a_i\right) = \begin{cases} -\dfrac{a_i^2 s}{2C^2}, & x_i \le s, \\ (1-2t)\dfrac{1}{2}s, & x_i > s, \end{cases} \qquad T\left(a_i^*\right) = \begin{cases} -\dfrac{a_i^* s}{2C^2}, & x_i^* \le s, \\ (1-2t)\dfrac{1}{2}s, & x_i^* > s. \end{cases}$$

Стоит отметить, что в двойственных задачах появляются дополнительные динамические ограничения, и требуется вносить изменения в классический алгоритм оптимизации алгоритма опорных векторов. Предлагаются два алгоритма решения подобных задач. Первый алгоритм решения полученной двойственной задачи (5), (6) можно сформулировать следующим образом.

Шаг 1. Найдем некоторое начальное решение $f(x)$, используя функцию потерь Лапласа.

Шаг 2. Вычислим значения x_i, x_i^* как расстояние от элементов выборки до полученного решения $x_i = \begin{cases} f(x_i) - y_i, & f(x_i) - y_i \geq 0 \\ 0, & f(x_i) - y_i < 0 \end{cases}$,

$$x_i^* = \begin{cases} y_i - f(x_i), & y_i - f(x_i) \geq 0 \\ 0, & y_i - f(x_i) < 0 \end{cases}, \; i = 1,...,n.$$

Шаг 3. Используя вычисленные значения x_i, x_i^*, зафиксируем ограничения (6) и решим задачу (5), используя, например, алгоритм последовательной оптимизации (SMO алгоритм).

Шаг 4. Вычислим значения x_i, x_i^* для нового решения. Если для полученных значений x_i, x_i^* ограничения (6) изменяются и на предыдущих шагах подобной их комбинации не встречалось, то переходим на шаг 3. В противном случае решение, полученное на шаге 3, является окончательным.

Для предотвращения возможного зацикливания в предложенной выше оптимизационной процедуре для всех точек выборки ведется учет типов ограничений, которые фиксируются на шаге 3, для двойственных переменных на каждой итерации алгоритма. При появлении уже встречавшейся ранее комбинации ограничений, работа алгоритма заканчивается. Очевидно, что в силу конечного числа возможных комбинаций ограничений, гарантируется завершение алгоритма. Многочисленные проведенные вычислительные эксперименты показывают, что алгоритм сходится за несколько шагов. Дополнительно для контроля достижения точки оптимума используется сравнение значений целевых функций прямой и двойственной задачи. В точке оптимума их значения должны совпадать.

Альтернативный алгоритм, основанный на использовании функции ε-нечувствительности Вапника, приводится в п. 2.3.4.

В п. 2.4 приводятся результаты исследований робастности алгоритма опорных векторов с различными функциями потерь на нескольких модельных примерах. Показано, что предложенные адаптивные функции потерь Лапласа и Хьюбера обеспечивают робастность в условиях не только «тяжелых хвостов», но и в условиях асимметричных засорений.

Глава 3. Конструирование двойственной задачи SVM с асимметричными функциями потерь

В данной главе приведена теоретическая основа восстановления зависимостей на основе SVM для случая ошибок наблюдений, имеющих асимметричное распределение.

Существует целый ряд задач, в которых распределение ошибок наблюдений не является симметричным. В п. 3.1 приводятся примеры подобных задач.

Можно выделить два основных класса асимметричных распределений: изначально асимметричные распределения (например, распределение экстремального значения) и скошенные распределения, полученные на основе базового симметричного распределения (скошенное Лапласа, Стьюдента и т.д.).

Опираясь на вид оптимальных функций потерь для различных плотностей распределений ошибок наблюдений, для практического использования можно конструировать их аппроксимации. При этом следует учитывать, что в алгоритме опорных векторов для поиска решения традиционно используется задача квадратичного программирования, что сужает класс используемых для аппроксимации функций до линейных и квадратичных. Получаемые таким способом аппроксимации функций потерь будем называть *квазиоптимальными*.

В п. 3.2 показывается построение квазиоптимальных функций потерь на основе линейно-квадратичных аппроксимаций для использования их в SVM. В случае несимметричного закона распределения ошибок наблюдений оптимизационную задачу в SVM можно сформулировать следующим образом

$$\min_{w,b,x,x^*}\left[\frac{1}{2}w^T w + C\sum_{k=1}^{n}\left(L(x_k)+L'(x_k^*)\right)\right]$$

при ограничениях

$$y_k - w^T j(x_k) - b \le e + x_k,$$
$$-y_k + w^T j(x_k) + b \le e + x_k^*,$$
$$x_k \ge 0,\ x_k^* \ge 0,\ k = 1,\mathbf{K},n,$$

где $L(x)$ и $L'(x)$ – функции потерь, используемые при отклонении наблюдений в ту или другую сторону от линии регрессии, $j(x)$ – используемое нелинейное отображение исходных данных в пространство большей размерности.

Возможны следующие варианты параметризации этих функций потерь:

1) $L(x) = qh(x)$, $L'(x) = (1-q)h(x)$;

2) $L(x) = h(qx)$, $L'(x) = h\left((1-q)x\right)$;

3) $L(x) = qh(x)$, $L'(x) = (1-q)h'(x)$.

Здесь $h(x)$ и $h'(x)$ – некоторые функции потерь, например, Гаусса, Лапласа или Хьюбера, а параметр $q \in (0,1)$ призван учитывать разницу в углах наклона правой и левой ветвей аппроксимируемой функции потерь.

В п. 3.3 приводится формулировка двойственной задачи для случая асимметричных функций потерь в SVM:

$$\max_{a_i,a_i^*} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_i - a_i^*\right)\left(a_j - a_j^*\right)j^T(x_i)j\left(x_j\right) + \sum_{i=1}^{n}\left(a_i - a_i^*\right)y_i$$
$$-e\sum_{i=1}^{n}\left(a_i + a_i^*\right) + C\sum_{i=1}^{n}\left(L(x_i) + L'\left(x_i^*\right) - x_i\frac{d}{dx_i}L(x_i) - x_i^*\frac{d}{dx_i^*}L'\left(x_i^*\right)\right) \tag{7}$$

Для функции потерь Лапласа $h(x) = |x|$, используя вариант параметризации №1, получаем

$$\max_{a_i, a_i^*} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_i - a_i^*\right)\left(a_j - a_j^*\right) j^T(x_i) j\left(x_j\right) + \sum_{i=1}^{n}\left(a_i - a_i^*\right) y_i - e\sum_{i=1}^{n}\left(a_i + a_i^*\right)$$

при ограничениях

$$\sum_{i=1}^{n}(a_i - a_i^*) = 0,\ a_i \in [0, Cq], a_i^* \in [0, C(1-q)]. \qquad (8)$$

Аналогично, подставляя в качестве $h(x)$ функцию потерь Хьюбера, для вариантов параметризации №1 и №2, соответственно, получаем следующий вид последнего слагаемого в (7):

1) $$CT(a) = -\frac{s}{2C}\left[\sum_{i=1}^{n}\left(\frac{a_i^2}{q} + \frac{a_i^{*2}}{1-q}\right)\right];$$

2) $$CT(a) = -\frac{s}{2C}\left[\sum_{i=1}^{n}\left(\frac{a_i^2}{q^2} + \frac{a_i^{*2}}{(1-q)^2}\right)\right].$$

Ограничения, накладываемые на переменные a_i и a_i^*, совпадают с (8).

Для варианта параметризации №3, когда левая ветвь аппроксимируется функцией потерь Лапласа, а правая – функцией потерь Хьюбера, т.е. $h(x) = |x|$ и $h'(x) = \begin{cases} \frac{1}{2s}x^2, & |x| \le s \\ |x| - \frac{s}{2}, & |x| > s \end{cases}$ получаем $CT(a) = -\frac{s}{2C}\left[\sum_{i=1}^{n}\frac{a_i^{*2}}{(1-q)^2}\right]$. Ограничения, накладываемые на переменные a_i и a_i^*, также совпадают с (8).

В п. 3.4 приводятся исследования алгоритма опорных векторов с различными функциями потерь в условиях асимметричных распределений ошибок наблюдений на основе нескольких модельных примеров. Показано, что использование асимметричных функций потерь позволяет существенно уменьшить среднеквадратичную ошибку аппроксимации в сравнении с симметричными функциями потерь.

В п. 3.5 рассматривается вопрос оценки параметра скошенности распределения q. Приводятся различные методы оценки в зависимости от степени априорной информации.

В п. 3.6 приводится формулировка квантильной регрессии на основе SVM с использованием асимметричных функций потерь. Дается двойственная формулировка задачи SVM для этого случая. Рассматриваются различные варианты учета асимметрии функции потерь.

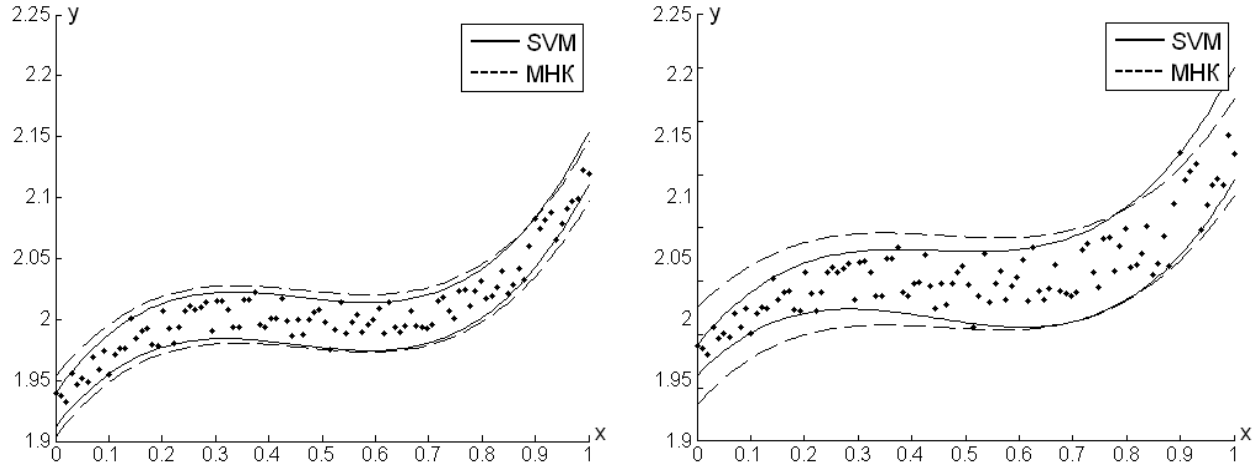
Двойственная задача для построения квантильной регрессии на основе SVM при использовании функции потерь Хьюбера принимает вид

$$\max_{a_i,a_i^*} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_i-a_i^*\right)\left(a_j-a_j^*\right)j^T\left(x_i\right)j\left(x_j\right)+\sum_{i=1}^{n}\left(a_i-a_i^*\right)y_i$$

$$-\frac{1}{2}\frac{s}{C}\sum_{i=1}^{n}\left[\frac{a_i}{q^2}+\frac{a_i^*}{(1-q)^2}\right]$$

при ограничениях (8). В данном случае $q \in (0,1)$ – заданная квантиль.

В п. 3.7 рассматривается вопрос построения доверительных интервалов на основе квантильной регрессии. В частности, задавая различные значения параметра q, можно строить доверительные интервалы для отклика. В условиях постоянства дисперсии ошибок наблюдений (рис. 2(а)), построенные на основе SVM интервалы оказываются близки к интервалам, построенным на основе классического подхода с использованием метода наименьших квадратов (МНК). При этом, в отличие от классического подхода, предложенный метод построения доверительного интервала для отклика можно использовать также и в условиях, когда дисперсия ошибок наблюдений не является постоянной (рис. 2(б)).



а) б)

Рис. 2. 95% доверительные интервалы для отклика когда: а) дисперсия ошибок постоянная; б) дисперсия ошибок переменная (линейно возрастает вдоль оси абсцисс)

В п. 3.8 показывается, как квантильная регрессия на основе SVM может быть использована для построения оценок неизвестной дисперсии ошибок наблюдений.

Глава 4. Построение разреженных решений

В данной главе приводятся методы построения разреженных решений при построении регрессионных моделей на основе SVM.

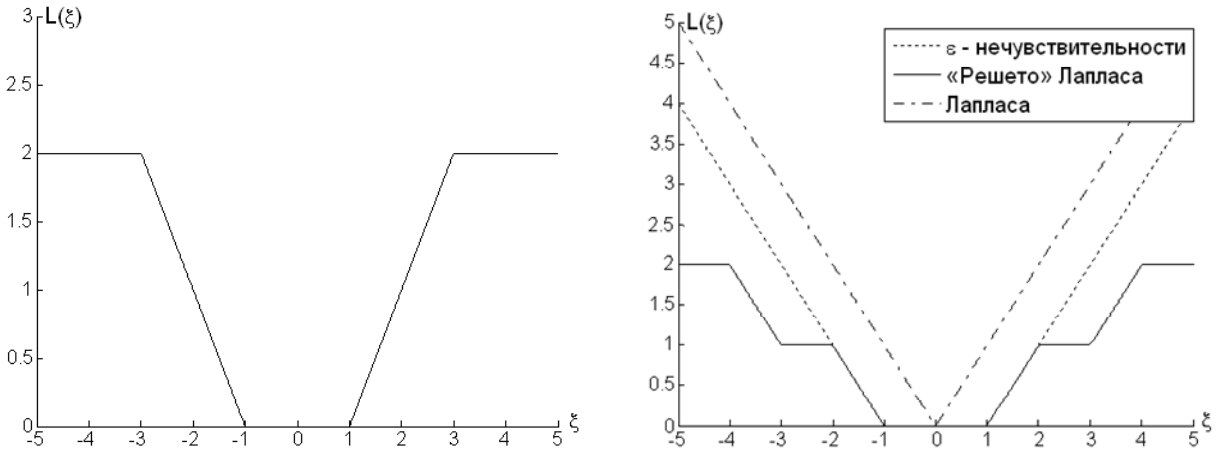
В п. 4.1 формулируется задача построения компактной модели регрессии. Приводится обзор подходов к решению данной задачи.

В п. 4.2 рассматривается механизм получения разреженных решений на основе функции ε-нечувствительности Вапника.

В п. 4.3 описывается использование предложенных адаптивных функций потерь для получения разреженных решений. Если говорить о разреженности в терминах функции потерь $L(x)$, или точнее относительно ее графика от аргумента x, то возникновение разреженных решений обусловлено наличием участков постоянства (зон нечувствительности). Решение формируется на опорных векторах x_i, не попадающих на участки постоянства функции $L(x)$. Недостатком функции потерь e-нечувствительности Вапника является то, что эти участки непостоянства при относительно широкой полосе e могут приходиться на хвосты распределения ошибок наблюдений.

Для устранения данного недостатка предлагается расширение функции e-нечувствительности Вапника на случай нескольких зон нечувствительности, которые располагаются на различном удалении от нулевой точки. Для этого предлагается использовать адаптивные функции потерь Лапласа и Хьюбера с параметром $t = 0$. На рис. 3(а) представлен график адаптивной функции потерь Лапласа.

Подход, описанный в п. 4.3, обобщается на случай произвольного числа зон нечувствительности в п. 4.4. Данный метод назван «решето» Лапласа. Функция потерь для этого метода показана на рис. 3(б).



а) б)

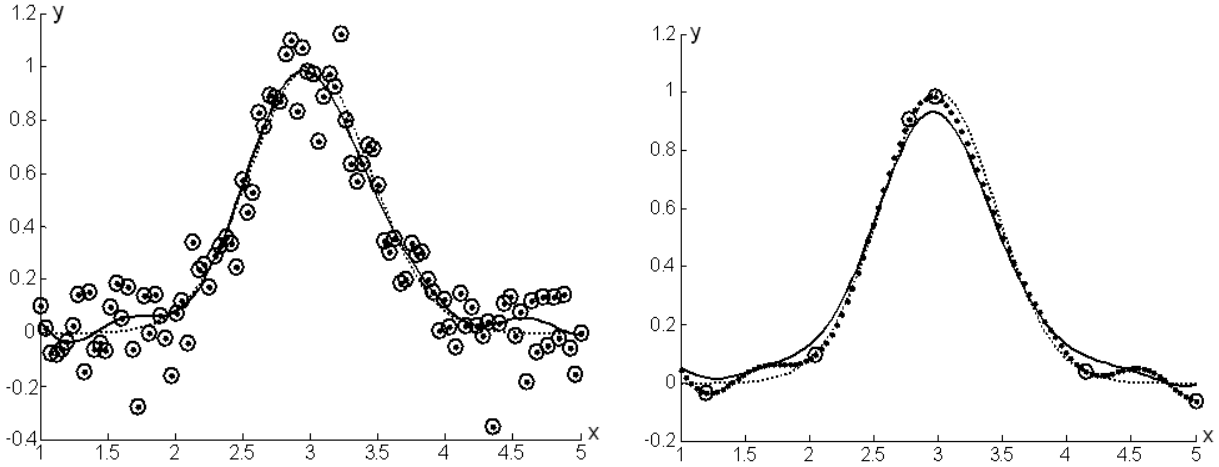
Рис. 3. Вид функций потерь для построения разреженных решений:
а) адаптивная функция потерь Лапласа при $\tau = 0$ ($e = 1,\ s = 3$);
б) функция потерь метода «решето» Лапласа

В п. 4.5 представлен другой метод получения разреженных решений – двухшаговый метод аппроксимации. Суть данного метода заключается в следующем. Сначала строится обычное неразреженное решение на основе какой-либо функции потерь, например, Лапласа или Хьюбера. Полученное решение должно удовлетворять исследователя по качеству аппроксимации, степени гладкости и другим необходимым свойствам. Данное базовое или исходное решение будет на втором шаге аппроксимировано разреженным решением, которое формируется при использовании функции потерь ε-нечувствительности Вапника. Для этого по полученной исходной модели генерируется необходимое число наблюдений, по возможности равномерно размещенных во всем про-

странстве определения исходных переменных. Эти наблюдения образуют множество *виртуальных опорных векторов,* на базе которых и будет построено разреженное решение.

Основную идею данного метода иллюстрирует рис. 4. Здесь окружностями обозначены опорные векторы, сплошной линией – истинная функция, пунктирной – SVM-регрессия. Очевидно, что в силу конструкции метода, наибольшая разреженность, при сохранении высокой точности, будет достигаться для слабо осциллирующих функций.

В п. 4.6 рассматривается проблема построения разреженных решений в условиях гетероскедастичности ошибок наблюдений. Для учета гетероскедастичности необходимо чтобы ширина зоны нечувствительности менялась в зависимости от величины дисперсии. Чем выше величина дисперсии, тем шире должна быть зона нечувствительности. На участках, где дисперсия мала, зона нечувствительности должна быть узкой. Показано, что благодаря учету гетероскедастичности, удается не только существенно уменьшить среднеквадратичную ошибку аппроксимации, но и получить при этом более разреженное решение.



а) б)

Рис. 4. Двухшаговый метод аппроксимации: а) начальное решение; б) аппроксимация решения на основе сгенерированных наблюдений

В п. 4.7 приводятся исследования различных методов построения разреженных решений. В качестве моделей, порождающих данные, использовались две функции: $r_1(x) = \exp\left(-(x-3)^2 / 0.4\right)$ и $r_2(x) = \sin(x)\cos(x^2)$. Результаты исследований представлены в таблице 1.

Исследования показали, что предложенные методы предоставляют возможность получения разреженных решений при сохранении высокой точности аппроксимации данных (низком значении среднеквадратичной ошибки аппроксимации – *MSE*). Их важным отличием от существующих методов является стабильность относительного числа опорных векторов (*S*) при увеличении объема выборки и степени зашумления данных.

Таблица 1

Используемая функция потерь / метод	*Уровень помехи*	$r_1(x)$		$r_2(x)$	
		MSE	*S*	*MSE*	*S*
Лапласа	10%	0.0489	100.0%	0.1066	100.0%
	20%	0.0680	100.0%	0.1473	100.0%
	30%	0.0964	100.0%	0.1683	100.0%
ε-нечувствительности Вапника	10%	0.0700	9.1%	0.1559	13.8%
	20%	0.0778	17.8%	0.1438	19.6%
	30%	0.0738	24.1%	0.1746	29.1%
Адаптивная функция потерь Лапласа	10%	0.0451	14.3%	0.1021	28.4%
	20%	0.0769	14.4%	0.1489	28.2%
	30%	0.0808	13.1%	0.1699	26.9%
Двухшаговый метод аппроксимации	10%	0.0519	10.6%	0.0889	23.7%
	20%	0.0586	11.1%	0.1331	25.4%
	30%	0.0815	12.0%	0.1801	25.4%

Глава 5. Применение SVM в задачах восстановления зависимостей

В данной главе исследованы возможности SVM для построения параметрических и полупараметрических моделей регрессии. Приведены модификации SVM для учета автокорреляции и гетероскедастичности ошибок наблюдений. На реальных данных продемонстрировано использование предложенных модификаций SVM для построения регрессионных моделей.

В п. 5.1 показана возможность построения параметрических моделей на основе SVM при использовании полиномиальных ядерных функций. Если использовать в SVM функцию потерь Лапласа, то в случае зашумления с тяжелыми хвостами (Лапласа, Коши), SVM существенно превосходит МНК как по точности оценок параметров, так и по значению среднеквадратичной ошибки аппроксимации. Очевидно, что если в SVM использовать функцию потерь Гаусса, то результаты окажутся близки к тем, что получаются в результате использования МНК. Таким образом, используя SVM для построения параметрических моделей, можно оценивать параметры моделей также как и в классических параметрических методах (например, МНК). Основным преимуществом SVM в данном случае можно считать возможность получения решений с использованием различных функций потерь и гарантию единственности решения.

В п. 5.2 приводится метод построения полупараметрических моделей на основе SVM. Предложенный подход базируется на использовании комбинации ядерных функций. Как известно, линейная комбинация ядер с положительными весами также является ядром. Благодаря этому свойству ядер можно комбинировать различные ядерные функции. Иногда на практике встречаются ситуации, когда в рассматриваемой зависимости явно прослеживается некий глобальный тренд (к примеру, в финансовых рядах, когда идет восходящий тренд с периодическими колебаниями). В этом случае целесообразно использовать смесь ядер: первое ядро будет описывать глобальный тренд, второе – локаль-

ные колебания. К примеру, если взять комбинацию полиномиального и Гауссова ядер:

$$K(x_i, x_j) = m \exp\left(-\frac{(x_i - x_j)^2}{2s^2} \right) + (1-m)(x_i x_j + 1)^d ,$$

где d – степень полинома, s – ширина ядра, $m \in [0,1]$ – параметр смеси, то получим возможность описывать основной тренд гладкими полиномами, а локальные осцилляции зависимости будут описываться с использованием Гауссова ядра.

В п. 5.3 представлены модификации SVM для построения регрессионных моделей в условиях гетероскедастичности ошибок наблюдений. Предложены три подхода для учета эффекта гетероскедастичности в SVM, два из которых позволяют использовать SVM в условиях отсутствия априорных знаний о характере изменения дисперсии ошибок наблюдений.

Первый подход основан на использовании ε-нечувствительной функции потерь с зоной нечувствительности, пропорциональной дисперсии наблюдений.

Второй подход, не связанный с параметризацией функции, описывающей поведение дисперсии отклика, состоит в следующем. Это обычная многошаговая схема (минимум два шага), когда на первом шаге оцениваются остатки, а на втором шаге идет окончательная (или промежуточная) оценка решения с учетом величин этих остатков. В этом случае весь интервал оцененных на первом шаге остатков разбивается на несколько равночастотных интервалов, и для наблюдений из этих интервалов назначается свой коридор нечувствительности. В этом случае влияние наблюдений с различной дисперсией уравнивается. Число интервалов разбиения может варьироваться. Чем выше темп изменения дисперсии, тем больше интервалов разбиения должно быть для более точного учета этих изменений. Данный подход можно считать непараметрическим вариантом учета гетероскедастичности.

Третий подход заключается в использовании оценок величины дисперсии на основе квантильной регрессии, которые были представлены в главе 3. Данный подход также можно считать непараметрическим, поскольку он не требует наличия априорной информации о характере изменения дисперсии ошибок наблюдений.

В п. 5.4 исследуется применение SVM в условиях мультиколлинеарности данных. Явление мультиколлинеарности возникает, если между объясняющими переменными существуют почти точные линейные зависимости (в интервале их изменения в эксперименте). В условиях мультиколлинеарности данных при использовании МНК приходится иметь дело с близкой к вырожденной матрицей $\Omega = Z^T Z$, где Z – матрица регрессоров. Одним из вариантов решения проблемы вырожденности матрицы Ω является использование регуляризации. К примеру, в методе ридж-оценок для улучшения обусловленности матрицы Ω к ней добавляется диагональная матрица.

На основе вычислительных экспериментов показано, что SVM нечувствителен к эффекту мультиколлинеарности данных. При использовании SVM в

этом случае все параметры, кроме тех, которые соответствуют «коррелированным» регрессорам, определяются достаточно точно. Оценки параметров модели, полученные с использованием SVM, оказываются близки к ридж-оценкам. Однако, в отличие от ридж-оценок, в SVM имеется возможность использования робастных функций потерь, таких как функции потерь Лапласа и Хьюбера.

В п. 5.5 исследуется применение SVM в условиях автокорреляции ошибок наблюдений. Предлагается модификация SVM для учета эффекта автокорреляции первого порядка. Показано, что при автокорреляции первого порядка с известным параметром автокорреляции r, в SVM необходимо выполнить замену переменных $y'_k = y_k - r\,y_{k-1}$, $k = \overline{2,n}$, $b' = (1-r)b$. Ядерную функцию, используемую при построении регрессионной модели, необходимо заменить на $K'(x_i, x_j) = K(x_i, x_j) - r\,K(x_{i-1}, x_j) - r\,K(x_i, x_{j-1}) + r^2 K(x_{i-1}, x_{j-1}), \qquad i, j = \overline{2,n}$. При этом отклик для модели будет вычисляться следующим образом: $y(x) = \sum_{i=2}^{n} a_i \left[K(x, x_i) - r\,K(x, x_{i-1})\right] + b'$.

В п. 5.6 рассматривается проблема выбора параметров метода SVM. Исследуется вопрос подбора оптимальных значений параметров алгоритма SVM. Показано, что эффективным вариантом выбора параметров является их подбор на основе вложенных сеток вокруг эвристически выбранных значений параметров.

В п. 5.7 исследуется применение метода SVM в прикладных задачах. В качестве первого примера рассматривается технологический процесс химического производства. Анализ этого процесса производится на основе трех различных откликов, которые принимают свои значения в зависимости от значений пяти факторов. Показано, что применение предложенных адаптивных функций потерь позволяет повысить качество регрессионной модели. В качестве других примеров используются широко распространенные выборки данных из сети интернет: «LIDAR», «Motorcycle», «Boston Housing». На примере этих выборок показана эффективность предложенных модификаций SVM в условиях гетероскедастичности ошибок наблюдений и наличия выбросов в данных.

Заключение

Основные результаты могут быть сформулированы следующим образом:

1. Для решения задачи устойчивого оценивания модели регрессии по технологии SVM в условиях зашумленных данных с помехой, имеющей распределение с «тяжелыми хвостами» или имеющей асимметричное засорение, предложено использование адаптивных функций потерь. Сформулирована двойственная задача для этого случая и реализована итерационная схема решения задачи квадратичного программирования с динамическими ограничениями.
2. Для построения регрессионных моделей в условиях, когда ошибки наблюдений имеют асимметричное распределение, предложено использование

асимметричных функций потерь в методе SVM. Сформулирована прямая и двойственная задачи для этого случая.

3. Обобщен метод квантильной регрессии на основе SVM на случай произвольной функции потерь. На его основе предложен метод построения доверительных интервалов для отклика, а также непараметрический метод оценки неизвестной дисперсии ошибок наблюдений.
4. Для построения компактной модели регрессии в условиях работы с выборками большого объема разработаны алгоритмы построения разреженных решений в SVM. Показана их эффективность в сравнении с классическим методом построения разреженных решений на основе функции нечувствительности Вапника. Предложена модификация SVM, позволяющая строить разреженные решения в условиях гетероскедастичности ошибок наблюдений.
5. Проведено экспериментальное исследование возможности построения регрессионных моделей с использованием SVM в условиях мультиколлинеарности данных, автокорреляции и гетероскедастичности ошибок наблюдений. Предложены модификации SVM для учета гетероскедастичности и автокорреляции ошибок наблюдений. Для предложенных модификаций сформулированы прямые и двойственные задачи SVM.
6. Разработана программная система для построения регрессионных моделей с использованием SVM. Разработанное программное обеспечение используется при проведении научных исследований.

Список публикаций

1. Попов А. А. Использование оценок степени гладкости функции при построении регрессии на основе метода опорных векторов / А. А. Попов, А. С. Саутин // Молодежь и современные информационные технологии : сб. тр. – Томск, 2008. – С. 149-150.
2. Попов А. А. Сравнение методов выбора параметров алгоритма опорных векторов в задаче построения регрессии / А. А. Попов, А. С. Саутин // Информатика и проблема телекоммуникаций: материалы российской науч.-технич. конф. – Новосибирск, 2008. – С. 74-77.
3. Саутин А. С. К вопросу о смещении решения в задаче построения регрессии с использованием алгоритма опорных векторов / А. С. Саутин // Современные информационные технологии : сб. статей. – Пенза, 2008. – С. 122-125.
4. Попов А. А. Анализ функций потерь в алгоритме опорных векторов при решении задачи построения регрессии / А. А. Попов, А. С. Саутин // Тр. междунар. конф. «Актуальные проблемы электронного приборостроения АПЭП-2008». – Новосибирск, 2008. – Т. 6. – С. 57-60.
5. Popov A. A. Selection of support vector machines parameters for regression using nested grids / A. A. Popov, A. S. Sautin // The Third International Forum on Strategic Technology. – Novosibirsk, 2008. – P. 329-331. [Выбор параметров алгоритма опорных векторов в задаче построения регрессионных моделей с использованием вложенных сеток]

6. Попов А. А. Определение параметров алгоритма опорных векторов при решении задачи построения регрессии / А. А. Попов, А. С. Саутин // Сб. научн. тр. НГТУ. – Новосибирск, 2008. – С. 35-40.
7. Popov A. A. Adaptive Huber Loss Function in Support Vector Regression / A. A. Popov, A. S. Sautin // The fourth international forum on strategic technology. – Hochiminh, Vietnam, 2009. – P. 114-118. [Адаптивная функция потерь Хьюбера в задаче построения регрессионных моделей на основе алгоритма опорных векторов]
8. Попов А. А. Построение регрессии по методу опорных векторов с ошибками наблюдений, имеющими асимметричное распределение / А. А. Попов, А. С. Саутин // Доклады АН ВШ РФ. – Новосибирск, 2009. – С. 117-126.
9. Попов А. А. Использование робастных функций потерь в алгоритме опорных векторов при решении задачи построения регрессии / А. А. Попов, А. С. Саутин // Научн. вестн. НГТУ. – 2009. – № 4(37). – С. 45-56. (из перечня ВАК)
10. Попов А. А. Построение разреженных решений при использовании алгоритма опорных векторов в задаче восстановления зависимости / А. А. Попов, А. С. Саутин // Научн. вестн. НГТУ. – 2010. – № 2(39). – С. 31-42. (из перечня ВАК)
11. Попов А. А. Оценивание дисперсии ошибок наблюдений с использованием квантильной регрессии на основе алгоритма опорных векторов / А. А. Попов, А. С. Саутин // Информатика и проблема телекоммуникаций: материалы российской науч.-технич. конф. – Новосибирск: Изд-во СибГУТИ, 2010. – Том I. – С. 90-93.