

На правах рукописи

ТАТАРЧУК Александр Игоревич

**БАЙЕСОВСКИЕ МЕТОДЫ ОПОРНЫХ ВЕКТОРОВ
ДЛЯ ОБУЧЕНИЯ РАСПОЗНАВАНИЮ ОБРАЗОВ
С УПРАВЛЯЕМОЙ СЕЛЕКТИВНОСТЬЮ ОТБОРА ПРИЗНАКОВ**

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук

Москва, 2014

Работа выполнена в Федеральном государственном бюджетном учреждении науки «Вычислительный центр им. А.А. Дородницына Российской академии наук».

Научный руководитель:

доктор технических наук, профессор **Вадим Вячеславович Моттль**

Официальные оппоненты:

Вьюгин Владимир Вячеславович, доктор физико-математических наук, профессор (Институт проблем передачи информации им. А. А. Харкевича РАН, заведующий лабораторией №1 им. М. С. Пинскера «Теория передачи информации и управления»)

Червоненкис Алексей Яковлевич, кандидат физико-математических наук, старший научный сотрудник (Федеральное государственное бюджетное учреждение науки «Институт проблем управления имени В.А. Трапезникова Российской академии наук», ведущий научный сотрудник)

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Московский государственный университет им. М.В. Ломоносова».

Защита диссертации состоится «19» июня 2014 г. в 13 ч. на заседании диссертационного совета Д 002.017.02 в ВЦ РАН по адресу 119991 Москва, ГСП-1, ул. Вавилова, 40.

С диссертацией можно ознакомиться в библиотеке и на официальном сайте (<http://www.ccas.ru/>) ВЦ РАН.

Автореферат разослан «__» _____ 2014 г.

Ученый секретарь

диссертационного совета Д 002.017.02

доктор физико-математических наук



Рязанов В.В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. В диссертации исследуется и развивается наиболее популярный в современной литературе линейный подход к обучению распознаванию двух классов объектов реального мира $\omega \in \Omega$, основанный на двух предположениях. Во-первых, предполагается, что всякий объект воспринимается компьютером через вектор его числовых признаков как точка в линейном пространстве $\mathbf{x}(\omega) \in \mathbb{R}^n$, размерность которого определяется числом признаков n . Во-вторых, предполагается, что для суждения о принадлежности объекта к одному из двух классов $y = \pm 1$ достаточно вычислить значение линейной решающей функции (decision function) $d(\mathbf{x} | \mathbf{a}, b) = \mathbf{a}^T \mathbf{x} + b: \mathbb{R}^n \rightarrow \mathbb{R}$, знак которой непосредственно укажет класс объекта $\mathbf{a}^T \mathbf{x} + b \gtrless 0$. Очевидно, что линейная функция $d(\mathbf{x} | \mathbf{a}, b)$ определяет дискриминантную гиперплоскость в $\mathbb{R}^n$, а решение о классе объекта определяется тем, по какую сторону от нее окажется вектор признаков объекта $\mathbf{x}$. Обучение линейного классификатора сводится к формированию значений параметров $(\mathbf{a}, b)$ на основе анализа конечной обучающей совокупности $\{(\mathbf{x}_j, y_j), j = 1, \dots, N\}$.

Наиболее популярным методом обучения линейного классификатора по обучающей совокупности является *метод опорных векторов* (Support Vector Machine – SVM), разработанный Вапником В.Н. и Червоненкисом А.Я.¹, в основе которого лежит ими же ранее предложенный метод обобщенного портрета². В основе метода лежит естественная идея выбирать ту гиперплоскость, которая в обучающей совокупности разделяет векторы признаков объектов разных классов с наибольшим зазором, дополнительно штрафуя возможные нарушения некоторыми объектами этого общего «идеального» требования. В данной диссертации исследуется и развивается именно метод опорных векторов.

Одно из основных преимуществ метода опорных векторов заключается в том, что как задача обучения, так и решающее правило классификации новых объектов определяются не самими векторами признаков отдельных объектов $\mathbf{x}(\omega) \in \mathbb{R}^n$, а только их попарными скалярными произведениями $(\mathbf{x}(\omega'))^T \mathbf{x}(\omega''): \Omega \times \Omega \rightarrow \mathbb{R}$. Из этого факта следует, что нет противоречия между традиционным признаковым способом погружения объектов распознавания в линейное пространство и безпризнаковым способом, предполагающим, что объекты $\omega \in \Omega$ могут быть представлены в компьютере только через некоторую числовую функцию парного отношения $K(\omega', \omega'')$. Правда, для того, чтобы идея поиска дискриминантной гиперплоскости в некотором линейном пространстве представления объектов реального мира имела простую математическую реализацию, функция $K(\omega', \omega''): \Omega \times \Omega \rightarrow \mathbb{R}$ должна обладать свойствами ядра (kernel function), т.е. быть симметричной и порождать неотрицательно определенные матрицы для всех конечных комбинаций объектов. Известно, что всякий ядро погружает множество объектов реального мира $\omega \in \Omega$ в некоторое большее гильбертово пространство $\tilde{\Omega} \supseteq \Omega$, в котором он играет роль скалярного произведения^{1,3}.

Однако требование неотрицательной определенности для функции парного сравнения объектов является чрезвычайно ограничивающим. Альтернативный подход предложен Р. Дьюиным⁴ под названием *реляционного дискриминантного анализа* (Relational Discriminant Analysis). Идея заключается в том, чтобы интерпретировать значения произвольной функции парного сравнения между всяким объектом $\omega \in \Omega$ и объектами из обучающей совокупности $\{\omega_1, \dots, \omega_N\}$ как вектор вторичных признаков этого объекта $\mathbf{x}(\omega) = (x_i(\omega) = K(\omega_i, \omega), i = 1, \dots, N)$, и применить затем обычный метод опорных векторов в $\mathbb{R}^N$. Однако за такое использование функций парного сравнения объектов приходится платить большой размерностью простран-

¹ Vapnik V. Statistical Learning Theory. NY.: J. Wiley, 1998.

² Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения). М.: Наука, 1974.

³ Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Метод потенциальных функций в теории обучения машин. М.: Наука, 1970.

⁴ R. Duin, E. Pekalska, D. Ridder. Relational discriminant analysis. Pattern Recognition Letters, Vol. 20, 1999, pp. 1175-1181.

ства вторичных признаков, определяемой числом объектов в обучающей совокупности. Эта размерность многократно увеличится, если использовать несколько альтернативных функций парного сравнения⁵.

С учетом последнего замечания в данной диссертации рассматривается метод опорных векторов в его традиционном признаковом понимании.

Первая проблемная ситуация, определившая выбор темы данного диссертационного исследования, заключается в том, что при всей эффективности метод опорных векторов остается эвристическим по своей конструкции. С момента его создания в мировой литературе был предпринят ряд попыток снабдить его некоторой вероятностной интерпретацией^{6,7}. Однако эти интерпретации оставались неполными, не позволяющими в полной мере использовать вероятностный аппарат для наделения чрезвычайно популярного метода опорных векторов принципиально новыми свойствами.

Для разрешения этой проблемной ситуации в настоящей диссертации предлагается специальная байесовская постановка обучения распознаванию двух классов объектов в линейном признаковом пространстве, приводящая к обобщению метода опорных векторов и являющаяся теоретической основой для создания новых селективных методов обучения. Основная идея байесовской постановки заключается в построении системы вероятностных предположений о паре плотностей распределения объектов двух классов $\phi(\mathbf{x}|y=\pm 1, \mathbf{a}, b)$, определяемой объективно существующей, но неизвестной гиперплоскостью $(\mathbf{a}, b)$ в линейном пространстве признаков, при некотором априорном предположении о ее случайном выборе $\Psi(\mathbf{a}, b)$. При этом именно структура первого семейства распределений определяет характерный принцип обучения по методу опорных векторов.

Что же касается априорного распределения параметров дискриминантной гиперплоскости, а именно ее направляющего вектора, то одна из основных идей диссертации заключается в том, что характер этого распределения определяет априорную склонность к увеличению одних компонент направляющего вектора и уменьшению других, предопределяя тем самым скрытое разделение признаков объектов на информативные и неинформативные. Выбор этого распределения играет роль регуляризации при формировании формальной байесовской постановки задачи обучения.

Вторая проблемная ситуация заключается в том, что избыточность признакового описания объектов неизбежно связана как с опасностью переобучения, выражающейся в снижении обобщающей способности результата обучения по ограниченной обучающей совокупности, так и с избыточной громоздкостью получаемых при этом решающих правил, что для некоторых практических задач является весьма обременительным.

Стремление избежать переобучения заставляет принимать специальные меры по ограничению свободы выбора решающего правила, называемые регуляризацией процесса обучения. Наиболее популярный принцип регуляризации обучения заключается в построении селективных систем распознавания, основанных на сокращении множества признаков, характеризующих объекты распознавания. В данной диссертации мы ограничимся рассмотрением именно методов отбора признаков объектов распознавания, независимо от того, каким способом эти признаки были получены.

Для разрешения второй проблемной ситуации в диссертации предлагается использовать выбор априорного распределения направляющего вектора для наделения метода опорных векторов в его вероятностной интерпретации способностью автоматически отбирать подмножество информативных признаков с заданным уровнем селективности в ре-

⁵ Середин О.С., Моттль В.В., Татарчук А.И., Разин Н.А. Выпуклые селективные критерии релевантных векторов для сокращения размерности описания объектов, представленных парными отношениями. Известия ТулГУ, Серия Естественные науки. Тула: Изд-во ТулГУ, 2013, Вып. 1, с. 165-176.

⁶ John C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in large margin classifiers, 1999, pp. 61-74, MIT Press.

⁷ Peter Sollich. Probabilistic methods for Support Vector Machines. Advances in Neural Information Processing Systems 12, 2000, pp. 349-355, MIT Press.

зультате обучения по предъявленной обучающей совокупности. Разнообразие существующих методов отбора признаков интерпретируется в диссертации как результат использования разных семейств априорных распределений направляющего вектора.

В диссертации впервые вводится понятие *параметра селективности*, являющегося параметром семейства априорных распределений направляющего вектора и выступающего в роли структурного параметра метода обучения. При нулевом значении параметра селективности оценка направляющего вектора содержит все исходные признаки объектов, а при его бесконечном увеличении все большая часть из них подавляется.

На основе анализа мировой литературы в диссертации приняты четыре критерия для оценивания «качества» селективных свойств конкретного метода обучения:

- а) эффективное подавление полностью шумовых попарно независимых признаков;
- б) эффективное подавление полностью шумовых признаков, имеющих значительную линейную зависимость между собой;
- в) эффективное выделение группы информативных линейно независимых признаков;
- г) эффективное выделение группы информативных признаков, только совместное участие которых в решении может обеспечить приемлемую точность распознавания.

По этим критериям в диссертации исследованы две наиболее популярных модификации метода опорных векторов, наделяющие его свойством отбора признаков – *Lasso SVM* (*1-norm SVM*) и *Elastic Net SVM* (*Doubly Regularized SVM*). Оба эти метода несколькими средствами отбирают подмножество информативных признаков, число которых определяется структурными параметрами.

Третья проблемная ситуация определяется тем, что, как оказалось, эти методы далеко не удовлетворяют сочетаниям пар требований (а-б) и (в-г).

Для разрешения этой проблемной ситуации в диссертации разработаны два новых класса априорных распределений направляющего вектора дискриминантной гиперплоскости и, следовательно, две новые модификации метода опорных векторов.

Один из новых методов, названный *методом релевантных признаков* (*Relevance Feature Machine – RFM*), не выделяя строгого подмножества информативных признаков, наделяет все признаки неотрицательными весами. Чем больше значение структурного параметра селективности, тем большее число весов приближаются к нулевым значениям, фактически исключая соответствующие признаки из принятия решения о классе объекта.

Другой предложенный метод, названный *методом опорных признаков* (*Support Feature Machine – SFM*), разбивает все множество признаков на три группы – полностью активные признаки, взвешенные признаки и полностью подавленные признаки. Можно считать, что метод *SFM* снабжает все признаки весами, но, в отличие от метода *RFM*, часть весов оказываются строгими единицами, часть принимает значения между единицей и нулем, а часть строго равны нулю.

Целью диссертации является разработка методологии селективного комбинирования признаков объектов в задаче двухклассового распознавания образов на основе байесовского обобщения метода опорных векторов.

Задачи исследования. Для достижения поставленной цели в диссертации сформулированы и решены следующие задачи:

1. Разработка общей математической постановки задачи обучения двухклассовому распознаванию образов в линейном пространстве признаков на основе количественного измерения расстояния между вектором признаков объекта и разделяющей гиперплоскостью.
2. Разработка вероятностной модели наблюдения объектов в пространстве векторов признаков относительно фиксированной разделяющей гиперплоскости, отражающей идею метода опорных векторов.
3. Разработка двух семейств априорных вероятностных моделей направляющего вектора разделяющей гиперплоскости, отражающих две принятые в мировой литературе стра-

тегии отбора признаков – взвешивания всех исходно заданных признаков (feature weighting) и жесткого выбора их подмножества (feature subset selection).

4. Разработка методов и алгоритмов байесовского оценивания разделяющей гиперплоскости, реализующих принцип обучения с заданной селективностью отбора признаков объектов.
5. Количественное экспериментальное исследование разработанных методов и алгоритмов на модельных данных и в ходе решения прикладных задач.

Методы исследования. Теоретическое исследование базируется на общих принципах линейной алгебры, методе опорных векторов, методах выпуклой оптимизации и основах байесовской теории принятия решений. Экспериментальное исследование проводилось с использованием программно-алгоритмического комплекса, разработанного автором.

Научная новизна. В данной работе впервые сформулирован вероятностный подход к селективному комбинированию признаков в ходе обучения распознаванию образов в рамках метода опорных векторов. Предложены два семейства параметрических вероятностных моделей обучающей совокупности и вытекающий из него класс линейных решающих правил и критериев обучения. Разработаны соответствующие алгоритмы апостериорного оценивания разделяющей гиперплоскости, реализующие байесовский принцип обучения с заданной селективностью отбора признаков объектов.

Положения, выносимые на защиту:

1. Общая математическая постановка задачи обучения двухклассовому распознаванию образов в линейном пространстве признаков на основе количественного измерения расстояния между вектором признаков объекта и разделяющей гиперплоскостью.
2. Вероятностная модель наблюдения объектов в пространстве векторов признаков относительно фиксированной разделяющей гиперплоскости.
3. Два семейства априорных вероятностных моделей направляющего вектора разделяющей гиперплоскости, отражающих стратегии отбора признаков на основе взвешивания всех исходно заданных признаков (feature weighting) и на основе жесткого выбора их подмножества (feature subset selection).
4. Комплекс байесовских методов и алгоритмов оценивания разделяющей гиперплоскости, реализующих принцип обучения с заданной селективностью отбора признаков объектов.

Практическая значимость. Разработанные алгоритмы позволяют строить решающие правила распознавания образов при заведомо избыточном множестве признаков представления объектов и относительно малом объеме обучающей совокупности без опасности снижения обобщающей способности результата обучения.

Связь с плановыми научными исследованиями. Работа выполнена в рамках гранта ИНТАС № 04-77-7347 «Принципы беспризнакового распознавания сигналов, символьных последовательностей и изображений» (2005-2006), грантов Российского фонда фундаментальных исследований № 05-01-00679-а «Линейные методы восстановления зависимостей в массивах данных произвольной природы», № 06-01-08042-офи «Теоретические основы, методы, инструментальные средства и новая открытая информационная технология построения систем идентификации личности по свободно пополняемому множеству биометрических характеристик», № 08-01-00695-а «Линейные методы комбинирования разнородной информации для решения задач анализа массивов данных произвольной природы», № 11-07-00409 «Методы выбора уровня сложности модели в задачах восстановления зависимостей между разнородными объектами реального мира».

Реализация и внедрение результатов работы. Результаты исследования применены для решения прикладных задач обучения распознаванию вторичной структуры белка по последовательности составляющих его аминокислот, распознаванию рукописных симво-

лов и подписей, вводимых в компьютер непосредственно в процессе написания, а так же в задаче распознавания рака легких.

Теоретические результаты диссертационной работы вошли в состав курса «Статистические методы анализа сигналов», читаемого профессором В.В. Моттлем студентам 5-го курса на кафедре «Интеллектуальные системы» ФУПМ МФТИ, и курсов «Теория обучения машин» и «Машинное обучение», читаемых профессором К.В. Воронцовым на ФУПМ МФТИ и в Школе анализа данных Яндекс, соответственно.

Апробация работы. Основные положения и результаты диссертации докладывались автором на конференциях:

- XIII всероссийская конференция «Математические методы распознавания образов», Зеленогорск, 2007 г.;
- The 7th International Workshop on Multiple Classifier Systems, Prague, Czech Republic, 2007 г.;
- The 6th International Conference on Machine Learning and Cybernetics, Китай, Гонконг, 2007 г.;
- Международная конференция «International Conference of Pattern Recognition», США, Тампа, 2008 г.;
- Международная конференция «Интеллектуализации обработки информации», Украина, Симферополь, 2008 г.;
- 14-ая Всероссийская конференция «Математические методы распознавания образов», Суздаль, 2009 г.;
- The 9th International Workshop, MCS 2010, Cairo, Egypt, 2010 г.;
- Международная конференция «International Conference of Pattern Recognition», Япония, Токио, 2012 г.

Кроме того, основные результаты работы докладывались на семинаре отдела Интеллектуальных систем (Москва, ВЦ РАН, 2009 г., 2014 г.).

Публикации. По тематике исследований опубликовано 18 статей, в том числе 8 статей в изданиях, входящих в список ВАК.

Структура работы. Диссертация состоит из введения, четырех глав, заключения, приложения и списка литературы. Материал изложен на 125 страницах, содержит 2 леммы, 15 теорем, 15 рисунков, 8 таблиц и список литературы из 21 наименования.

Благодарность. Автор глубоко признателен своему научному руководителю, учителю и наставнику профессору, д.т.н. Вадиму Вячеславовичу Моттлю за неоценимый вклад в научное мировоззрение автора. Автор также благодарен сотрудникам отдела «Интеллектуальные системы» ВЦ РАН и сотрудникам кафедры АТМ ТулГУ за внимание, помощь и поддержку в годы работы автора над диссертацией.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность исследований по разработке методов управления селективностью отбора признаков объектов в алгоритмах обучения распознаванию образов, сформулированы цели и задачи проводимых исследований, положения, выносимые на защиту, приведены сведения о структуре диссертации и ее апробации.

В первой главе дана краткая характеристика линейного подхода к обучению распознаванию двух классов объектов в некотором множестве объектов реального мира $\omega \in \Omega$, $y(\omega) = \pm 1$, которые воспринимаются компьютером через векторы их числовых признаков $\mathbf{x}(\omega) = (x_1(\omega), \dots, x_n(\omega)) \in \mathbb{R}^n$. Основополагающим принципом машинного обучения является *гипотеза компактности*, которая, в самой общей формулировке для задач обучения распознаванию образов, означает, что объекты одного и того же класса имеют большее сходство между собой чем объекты, принадлежащие разным классам. Самым распространённым математическим инструментом измерения *несходства* между объектами реального

мира Ω , погруженными числовыми признаками $\mathbf{x}(\omega) \in \mathbb{R}^n$ в конечномерное евклидовое пространство, является евклидова метрика $\rho(\omega', \omega'') = \rho(\mathbf{x}', \mathbf{x}'') = \|\mathbf{x}' - \mathbf{x}''\| = [(\mathbf{x}' - \mathbf{x}'')^T (\mathbf{x}' - \mathbf{x}'')]^{1/2}$.

Суть линейного подхода заключается в поиске такой гиперплоскости, которая обеспечивала бы в некотором смысле «хорошее» разделение объектов первого $y(\omega) = +1$ и второго $y(\omega) = -1$ классов. При этом всякий выбор направляющего вектора $\mathbf{a} \in \mathbb{R}^n$ и действительного числа $b \in \mathbb{R}$ определяет некоторую *разделяющую гиперплоскость* $\mathcal{H}(\mathbf{a}, b) = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{a}^T \mathbf{z} + b = 0\} \subset \mathbb{R}^n$. Ключевым понятием линейного подхода является *решающая функция* (decision score function), понимаемая как расстояние от объекта $\mathbf{x}(\omega) \in \mathbb{R}^n$ до его проекции $\mathbf{x}_{\mathcal{H}}(\omega) \in \mathbb{R}^n$ на гиперплоскость $\mathcal{H}(\mathbf{a}, b)$ с учетом знака в смысле евклидовой метрики:

$$d(\mathbf{x} | \mathbf{a}, b) = \frac{\mathbf{a}^T \mathbf{x} + b}{(\mathbf{a}^T \mathbf{a})^{1/2}} \begin{cases} > 0 \Rightarrow \hat{y} = 1, \\ < 0 \Rightarrow \hat{y} = -1, \end{cases} \text{ или } d(\mathbf{x} | \mathbf{a}, b) = \mathbf{a}^T \mathbf{x} + b \geq 0 \text{ при } \|\mathbf{a}\| = 1. \quad (1)$$

Пусть задана обучающая совокупность объектов с указанием индексов классов, к которым они принадлежат:

$$\{(\mathbf{x}_j, y_j), j = 1, \dots, N\}, \mathbf{x}_j \in \mathbb{R}^n, y_j = \pm 1. \quad (2)$$

Степень несоответствия значения решающей функции $d(\mathbf{x} | \mathbf{a}, b)$ для объекта $\mathbf{x}(\omega) \in \mathbb{R}^n$ значению его целевой характеристики $y(\omega)$ выражается в виде *функции потерь*. В частности, нас далее будет интересовать функция потерь специального вида (рис. 1)

$$\delta(y, d | \varepsilon) = \begin{cases} 0, & yd \geq \varepsilon, \\ 1 - \frac{1}{\varepsilon} yd, & yd < \varepsilon, \end{cases} \quad (3)$$

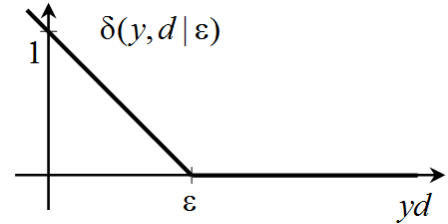


Рис. 1. Функция потерь $\delta(y, d | \varepsilon)$.

где $y = y(\omega)$ — истинный класс объекта, $d = d(\mathbf{x} | \mathbf{a}, b)$ — значение решающей функции (1), а $\varepsilon \geq 0$ — параметр, который далее будем интерпретировать как *зазор* разделения двух классов объектов.

Функция потерь такого вида была впервые предложена Вапником В.Н. и Червоненкисом А.Я. и привела к методу обобщенного портрета⁸, а затем к известному методу опорных векторов⁹. Такая функция потерь имеет нулевой штраф $\delta(y, d | \varepsilon) = 0$ для объектов, удаленных от гиперплоскости на «достаточное» расстояние $yd \geq \varepsilon$, и линейно растущий штраф $\delta(y, d | \varepsilon) = 1 - yd/\varepsilon$ для объектов, слишком близко приблизившихся к границе классов или попавших в область другого класса $yd < \varepsilon$.

В качестве критерия обучения естественно искать такую гиперплоскость, которая разделяла бы обучающую выборку на два класса, с одной стороны, с как можно большей величиной зазора $\varepsilon \geq 0$, $1/\varepsilon^2 \rightarrow \min$, а с другой, с как можно меньшей величиной суммарного штрафа для ошибочно классифицированных объектов обучающей выборки (2) $\delta_j = 1 - (1/\varepsilon)y_j(\mathbf{a}^T \mathbf{x}_j + b)$, $\sum_{j: (1/\varepsilon)y_j(\mathbf{a}^T \mathbf{x}_j + b) < 1} \delta_j$. Баланс таких требований к процессу обучения выражается оптимизационным критерием

$$\begin{cases} \varepsilon^{-2} + C \sum_{j=1}^N \delta_j \rightarrow \min(\mathbf{a}, b, \delta, \varepsilon), & \mathbf{a}^T \mathbf{a} = 1, \\ y_j(\mathbf{a}^T \mathbf{x}_j + b) \geq \varepsilon(1 - \delta_j), & \delta_j \geq 0, j = 1, \dots, N, \varepsilon \geq 0. \end{cases} \quad (4)$$

Здесь C — структурный параметр, определяющий соотношение требований максимизации зазора и минимизации величины эмпирического риска. Это и есть исходный критерий обучения, получивший название метода опорных векторов.

⁸ Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения). М.: Наука, 1974.

⁹ Vapnik V. Statistical Learning Theory. NY.: J. Wiley, 1998.

Правда, критерий (4) не очень удобен, поскольку имеет вид задачи оптимизации на сфере $\mathbf{a}^T \mathbf{a} = 1$, представляющей собой невыпуклое множество. Классический критерий получается из (4) простой заменой переменных $\mathbf{a} = \varepsilon \tilde{\mathbf{a}}$, $b = \varepsilon \tilde{b}$, $\tilde{\mathbf{a}}^T \tilde{\mathbf{a}} = 1/\varepsilon^2$:

$$\begin{cases} J_{SVM}(\mathbf{a}, b, \delta | C) = \mathbf{a}^T \mathbf{a} + C \sum_{j=1}^N \delta_j \rightarrow \min(\mathbf{a}, b, \delta), \\ y_j(\mathbf{a}^T \mathbf{x}_j + b) \geq 1 - \delta_j, \delta_j \geq 0, j=1, \dots, N. \end{cases} \quad (5)$$

Это задача квадратичного программирования – минимизации квадратичной функции при линейных ограничениях типа неравенств. Название метода опорных векторов (Support Vector Machine) объясняется тем, что направляющий вектор оптимальной разделяющей гиперплоскости $(\hat{\mathbf{a}}, \hat{b}, \hat{\delta}) = \arg \min J_{SVM}(\mathbf{a}, b, \delta | C)$ согласно (5) есть линейная комбинация векторов признаков лишь небольшого числа объектов обучающей совокупности⁹ $\hat{\mathbf{a}} = \sum_{j: \hat{\lambda}_j > 0} y_j \hat{\lambda}_j \mathbf{x}_j \in \mathbb{R}^n$, а именно только тех объектов, для которых в точке минимума критерия (5) ограничения $y_j(\mathbf{a}^T \mathbf{x}_j + b) \geq 1 - \delta_j$ являются активными, т.е. множители Лагранжа при них $(\hat{\lambda}_j \geq 0, j=1, \dots, N)$ строго положительны $\hat{\lambda}_j > 0$. Значения множителей Лагранжа дает двойственная задача квадратичного программирования⁹, решение которой эквивалентно решению исходной задачи (5). Вся решающая функция

$$d(\mathbf{x}) = \sum_{j: \hat{\lambda}_j > 0} \hat{\lambda}_j y_j \mathbf{x}_j^T \mathbf{x} + \hat{b} \geq 0, \quad \hat{b} = - \frac{\sum_{j: 0 < \lambda_j < C/2} \hat{\lambda}_j \sum_{l: \lambda_l > 0} y_l \hat{\lambda}_l \mathbf{x}_l^T \mathbf{x}_j + (C/2) \sum_{j: \lambda_j = C/2} y_j}{\sum_{j: 0 < \lambda_j < C/2} \hat{\lambda}_j}, \quad (6)$$

применяемая к вектору признаков нового объекта $\mathbf{x} \in \mathbb{R}^n$, полностью определяется лишь опорными объектами обучающей совокупности, для которых множители Лагранжа строго положительны $\hat{\lambda}_j > 0$.

В первой главе сформулирована основная идея селективного обучения, опирающаяся на тот факт¹⁰, что для оптимальной гиперплоскости, разделяющей без ошибок заданную обучающую совокупность размера N на два класса с зазором $\varepsilon > 0$, справедлива следующая оценка математического ожидания вероятности ошибки на генеральной совокупности (в среднем по всем обучающим совокупностям заданного размера)

$$EP_{error} \leq E \min \left(\frac{m}{N}, \frac{[R^2 \varepsilon^{-2}]}{N}, \frac{n}{N} \right), \quad (7)$$

Выражение (7) наглядно показывает три основных механизма, объясняющих хорошую обобщающую способность методов распознавания образов, основанных на понятии оптимальной разделяющей гиперплоскости:

- представление гиперплоскости небольшим количеством m опорных векторов;
- разделение выборки с наибольшим зазором ε между объектами разных классов;
- обучение в признаковом пространстве небольшой размерности n .

Классическая постановка метода опорных векторов (5) полностью игнорирует третий механизм повышения обобщающей способности обучения, состоящий в сокращении размерности признакового пространства.

В первой главе диссертации выполнен обзор и анализ существующих селективных модификаций метода опорных векторов, указана их недостаточность и **сформулированы идеи двух вероятностных модификаций метода опорных векторов**, а именно **метода релевантных признаков с регулируемой селективностью** (Relevance Feature Machine) и **метода опорных признаков с регулируемой селективностью** (Support Feature Machine), которые являются представителями, соответственно, непрерывных и дискретных встроженных

¹⁰ Vapnik V. N. The Nature of Statistical Learning Theory. Springer, 1995.

методов (Embedded Methods) отбора исходных представлений объектов, т.е. методов, сохраняющих все признаки объектов и снабжающих их лишь разными по величине положительными весами, в отличие от методов, жестко отбирающих подмножество признаков, называемых автором как опорные, при полном игнорировании остальных.

Во второй главе излагается общая вероятностная байесовская постановка задачи обучения двухклассовому распознаванию образов в линейном пространстве $\mathbf{x} \in \mathbb{R}^n$ на основе метода опорных векторов, сформулированного в его классическом виде⁹ в сугубо детерминистских терминах. Основная идея предлагаемой байесовской постановки заключается в построении специальной системы вероятностных предположений о паре плотностей распределения векторов признаков объектов двух классов $\varphi(\mathbf{x}|y=1)$ и $\varphi(\mathbf{x}|y=-1)$, определяемой объективно существующей, но неизвестной гиперплоскостью в линейном пространстве признаков, при некотором априорном предположении о ее случайном выборе.

Пусть $\mathbf{a}^T \mathbf{x} + b \geq 0$ – некоторая гиперплоскость с направляющим вектором $\mathbf{a} \in \mathbb{R}^n$ и параметром сдвига $b \in \mathbb{R}$. Свяжем с этой гиперплоскостью пару параметрических семейств условных плотностей распределения вероятностей вектора признаков случайного объекта

$$\varphi(\mathbf{x}|\mathbf{a}, b, y; c) = \text{const} \begin{cases} 1, & y(\mathbf{a}^T \mathbf{x} + b) \geq 1, \\ \exp[-c(1 - y(\mathbf{a}^T \mathbf{x} + b))], & y(\mathbf{a}^T \mathbf{x} + b) < 1. \end{cases} \quad (8)$$

Эти два семейства призваны выражать предположение, что случайные векторы признаков объектов двух классов распределены главным образом в «своих» полупространствах $\mathbf{a}^T \mathbf{x} + b > 0$ и $\mathbf{a}^T \mathbf{x} + b < 0$, однако могут попадать и в «чужие» полупространства, причем степенью возможности «ошибочных» значений управляет параметр $c > 0$. Некорректность равномерного распределения в бесконечной области в верхней строке (8) не приводит далее к математическим противоречиям, поскольку оно участвует только в формуле Байеса¹¹.

Будем предполагать, что получена обучающая совокупность (2), образованная векторами признаков объектов $\mathbf{x}_j = (x_{ij}, i=1, \dots, n)$ с известными индексами классов y_j . Тогда условное распределение обучающей совокупности в целом $(X|Y)$ представимо в виде произведения плотностей (8):

$$\Phi(X|Y, \mathbf{a}, b; c) = \prod_{j=1}^N \varphi(\mathbf{x}_j | \mathbf{a}, b, y_j; c). \quad (9)$$

Другим ключевым предположением в предлагаемой вероятностной модели является суждение об априорном совместном распределении $\Psi(\mathbf{a}, b)$ параметров $(\mathbf{a}, b)$ разделяющей гиперплоскости $\mathbf{a}^T \mathbf{x} + b \geq 0$. Будем считать, что отсутствуют какие-либо априорные предпочтения величин порога разделяющей гиперплоскости $b \in \mathbb{R}$, тогда

$$\Psi(\mathbf{a}, b) \propto \Psi(\mathbf{a}). \quad (10)$$

Апостериорная плотность распределения $P(\mathbf{a}, b|X, Y; c)$ параметров $(\mathbf{a}, b)$ относительно обучающей совокупности (X, Y) определяется формулой Байеса:

$$P(\mathbf{a}, b|X, Y; c) = \frac{\Psi(\mathbf{a}, b)\Phi(X|Y, \mathbf{a}, b; c)}{\text{const}} \propto \Psi(\mathbf{a}, b)\Phi(X|Y, \mathbf{a}, b). \quad (11)$$

Понимание обучения как задачи максимизации этой апостериорной плотности распределения в пространстве параметров модели $(\mathbf{a}, b)$ приводит к очевидному критерию:

$$(\hat{\mathbf{a}}, \hat{b}|X, Y; c) = \arg \max_{\mathbf{a} \in \mathbb{R}^n, b \in \mathbb{R}} \Psi(\mathbf{a}, b)\Phi(X|Y, \mathbf{a}, b; c) = \arg \max_{\mathbf{a} \in \mathbb{R}^n, b \in \mathbb{R}} (\ln \Psi(\mathbf{a}, b) + \ln \Phi(X|Y, \mathbf{a}, b; c)). \quad (12)$$

Байесовский критерий обучения (12) со структурой условного распределения наблюдений (8)-(9) является **основным теоретическим предложением данной диссертации**.

¹¹ ДеГроот М. Оптимальные статистические решения. М.: Мир, 1974.

Следующая теорема, доказанная в диссертации, показывает, что этот критерий представляет собой обобщение классического метода опорных векторов.

Теорема 1. *Критерий обучения (12) для семейства распределений (8) и априорного распределения параметров (10) эквивалентен оптимизационной задаче минимизации целевой функции $J(\mathbf{a}, \mathbf{b}, \boldsymbol{\delta} | c)$ на выпуклом множестве, заданном набором линейных ограничений-неравенств для объектов обучающей совокупности:*

$$\begin{cases} J(\mathbf{a}, \mathbf{b}, \boldsymbol{\delta} | c) = -\ln \Psi(\mathbf{a}) + c \sum_{j=1}^N \delta_j \rightarrow \min(\mathbf{a}, \mathbf{b}, \boldsymbol{\delta}), \\ y_j(\mathbf{a}^T \mathbf{x}_j + b) \geq 1 - \delta_j, \delta_j \geq 0, j = 1, \dots, N. \end{cases} \quad (13)$$

По своей структуре этот критерий отличается от классического критерия опорных векторов (5) только слагаемым $-\ln \Psi(\mathbf{a})$ вместо квадрата нормы направляющего вектора искомой разделяющей гиперплоскости $\mathbf{a}^T \mathbf{a}$. Параметр $c > 0$ семейств распределений (8), отвечающий за априорную способность объектов генеральной совокупности нарушать границу своих классов, регламентирует приоритет между качеством разделения обучающей совокупности и априорными предпочтениями по выбору направляющего вектора.

Если функция $\ln \Psi(\mathbf{a})$ строго вогнута, то целевая функция $J(\mathbf{a}, \mathbf{b}, \boldsymbol{\delta} | c)$ в (13) строго выпукла. Поскольку система ограничений в (13) образует выпуклый многогранник в линейном пространстве $(\mathbf{a}, \mathbf{b}, \boldsymbol{\delta}) \in \mathbb{R}^n \times \mathbb{R}^{N+1}$ предлагаемый в диссертации обобщенный критерий обучения по методу опорных векторов в общем виде представляет собой задачу выпуклого программирования. Это значит, что существует единственный набор оптимальных параметров $(\hat{\mathbf{a}}, \hat{b})$, удовлетворяющий всем ограничениям задачи (13) и доставляющий минимум целевой функции $J(\mathbf{a}, \mathbf{b}, \boldsymbol{\delta} | c)$.

Результатом обучения по обучающей совокупности (2) является решающее правило классификации $d(\mathbf{x} | \hat{\mathbf{a}}, \hat{b}) = \hat{\mathbf{a}}^T \mathbf{x} + \hat{b} \geq 0$, применимое к любому новому объекту с вектором признаков $\mathbf{x} \in \mathbb{R}^n$.

Принципиальной новизной предлагаемой формулировки задачи обучения является тот факт, что предположение об условном распределении векторов признаков случайных объектов (8), выражающее основную специфику метода опорных векторов, позволяет интерпретировать полученное решающее правило не только как суждение о классе нового объекта, но и наделить это суждение естественной оценкой его «уверенности».

Действительно, примем предположение, что случайно появившийся новый объект, не содержащийся в обучающей совокупности, объективно принадлежит тому или иному классу с некоторыми априорными вероятностями $q_{-1} = P(y = -1)$, $q_1 = P(y = 1)$, $q_{-1} + q_1 = 1$. Для простоты дальнейших рассуждений примем, в частности, равными априорные вероятности двух классов $q_{-1} = q_1 = 1/2$ ¹². Тогда, для параметров разделяющей гиперплоскости $(\hat{\mathbf{a}}, \hat{b})$, найденных как решение задачи обучения (13), апостериорная вероятность принадлежности объекта $\mathbf{x} \in \mathbb{R}^n$ к классу $y = 1$ определяется формулой Байеса через плотности распределения двух классов $\varphi(\mathbf{x} | \hat{\mathbf{a}}, \hat{b}, y; c)$:

$$p(y = 1 | \mathbf{x}; \hat{\mathbf{a}}, \hat{b}, c) = 1 - p(y = -1 | \mathbf{x}; \hat{\mathbf{a}}, \hat{b}, c) = \frac{\varphi(\mathbf{x} | y = 1; \hat{\mathbf{a}}, \hat{b}, c)}{\varphi(\mathbf{x} | y = 1; \hat{\mathbf{a}}, \hat{b}, c) + \varphi(\mathbf{x} | y = -1; \hat{\mathbf{a}}, \hat{b}, c)}. \quad (14)$$

¹² Выбор разных априорных вероятностей классов $q_1 \neq q_{-1}$ отражает другое предположение об их свойстве в сравнении с выбором значения параметра пересечения классов $c > 0$, характеризующих их разнотональность. В данной работе роль априорных вероятностей классов (q_1, q_{-1}) как параметров вероятностной модели детально не исследуется.

Теорема 2. Для условных плотностей распределения (8) и для случая равных априорных вероятностей $q_{-1} = q_1 = 1/2$ апостериорные вероятности классов (14) будут иметь вид сигмоидоподобной функции

$$p(y=1|\mathbf{x}, \hat{\mathbf{a}}, \hat{b}; c) = 1 - p(y=-1|\mathbf{x}, \hat{\mathbf{a}}, \hat{b}; c) = \begin{cases} \left\{1 + \exp(c) \exp[-c(\hat{\mathbf{a}}^T \mathbf{x} + \hat{b})]\right\}^{-1}, & \hat{\mathbf{a}}^T \mathbf{x} + \hat{b} < -1, \\ \left\{1 + \exp[-2c(\hat{\mathbf{a}}^T \mathbf{x} + \hat{b})]\right\}^{-1}, & -1 \leq \hat{\mathbf{a}}^T \mathbf{x} + \hat{b} \leq 1, \\ \left\{1 + \exp(-c) \exp[-c(\hat{\mathbf{a}}^T \mathbf{x} + \hat{b})]\right\}^{-1}, & \hat{\mathbf{a}}^T \mathbf{x} + \hat{b} > 1. \end{cases} \quad (15)$$

Заметим, что параметр $c > 0$ условных распределений (8), определяющий степень пересечения классов, управляет крутизной сигмоидоподобного обобщения (15) ступенчатой решающей функции $\hat{\mathbf{a}}^T \mathbf{x} + \hat{b} \geq 0$. Соответствующая зависимость апостериорной вероятности (15) от значения решающей функции $d(\mathbf{x} | \hat{\mathbf{a}}, \hat{b}) = \hat{\mathbf{a}}^T \mathbf{x} + \hat{b}$ показана на рис. 2.

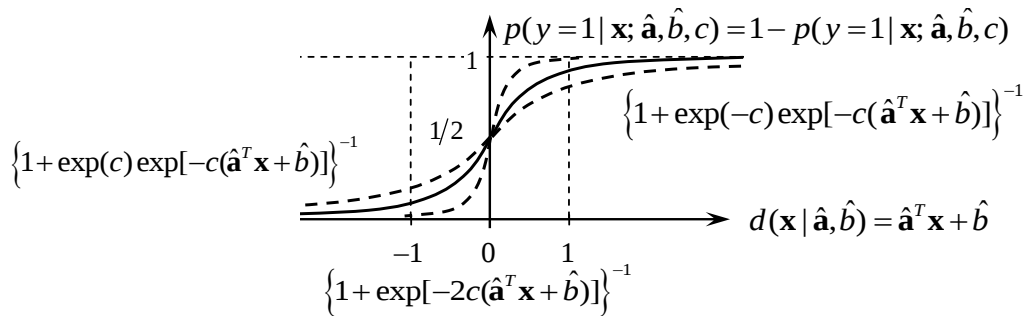


Рис. 2. Апостериорные вероятности классов для вновь поступившего объекта в вероятностной интерпретации метода опорных векторов.

Классический метод опорных векторов: Частный случай нормальных априорных распределений компонент направляющего вектора $\mathbf{a} \in \mathbb{R}^n$ с одинаковыми дисперсиями. Рассмотрим задачу обучения (13), приняв предположение, что случайный направляющий вектор образован независимыми компонентами $\mathbf{a} = (a_1, \dots, a_n)$ с нормальными плотностями распределения, имеющими нулевое математическое ожидание и одинаковые дисперсии $r_1 = \dots = r_n = r$. Тогда $\Psi(\mathbf{a} | r) = \left[1 / (r^{n/2} (2\pi)^n)\right] \exp\left(-(1/2r) \sum_{i=1}^n a_i^2\right)$, и задача обучения (13) примет вид классического критерия обучения по методу опорных векторов (5), где $C = 2cr$.

Заметим, что в диссертации впервые предложена формальная интерпретация структурного параметра C в методе опорных векторов как относительной степени пересечения классов по сравнению с общей априорной дисперсией компонент направляющего вектора $C = 2cr$. В такой интерпретации нет смысла рассматривать иные значения априорной дисперсии, нежели $r = 1/2$, тогда $C = c$.

Известные методы автоматического сокращения размерности признакового описания объектов в рамках метода опорных векторов, а именно, 1-norm SVM (Lasso SVM)¹³ и Doubly Regularized SVM (Elastic Net SVM)¹⁴ получаются из обобщенного критерия (13) специальным выбором априорной плотности распределения направляющего вектора $\Psi(\mathbf{a})$. Метод Elastic Net SVM, как более общий из них, получается из (13) при предположении, что случайный направляющий вектор образован независимыми компонентами $\mathbf{a} = (a_1, \dots, a_n)$ с одинаковыми плотностями распределения $\psi(a_i | \beta, \mu) = D \exp[-(\beta a_i^2 + \mu |a_i|)]$, которые определяются одинаковыми значениями параметров (β, μ) . Здесь нормирующая константа определяется обоими параметрами:

¹³ Bradley P., Mangasarian O. Feature selection via concave minimization and support vector machines. In International Conference on Machine Learning. 1998.

¹⁴ Li Wang, Ji Zhu, Hui Zou. The doubly regularized support vector machine. Statistica Sinica, 01/2006; 16:589-615.

$$D = \left[2\sqrt{\frac{\pi}{\beta}} \exp\left(\frac{\mu^2}{4\beta}\right) \Phi\left(\frac{\mu}{\sqrt{2\beta}}\right) \right]^{-1}, \text{ где } \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_u^{\infty} \exp\left(-\frac{z^2}{2}\right) dz - \text{функция Лапласа}^{15}.$$

В этом случае совместная априорная плотность распределения направляющего вектора определяется выражением

$$\Psi(\mathbf{a}|\beta, \mu) = D^n \exp\left[-\sum_{i=1}^n (\beta a_i^2 + \mu |a_i|)\right], \ln \Psi(\mathbf{a}|\beta, \mu) = \text{const} - \left(\beta \sum_{i=1}^n a_i^2 + \mu \sum_{i=1}^n |a_i|\right),$$

и (13) примет вид, эквивалентный методу Elastic Net SVM:

$$\begin{cases} J_{DrSVM}(\mathbf{a}, b, \delta | c, \beta, \mu) = \beta \sum_{i=1}^n a_i^2 + \mu \sum_{i=1}^n |a_i| + c \sum_{j=1}^N \delta_j \rightarrow \min(\mathbf{a}, b, \delta), \\ y_j \left(\sum_{i=1}^n a_i x_{ij} + b \right) \geq 1 - \delta_j, \delta_j \geq 0, j = 1, \dots, N. \end{cases} \quad (16)$$

В еще более частном случае, когда $\beta = 0$, получается метод обучения Lasso SVM:

$$\begin{cases} J_{1nSVM}(\mathbf{a}, b, \delta | c, r) = \sum_{i=1}^n |a_i| + c(r/2)^{1/2} \sum_{j=1}^N \delta_j \rightarrow \min(\mathbf{a}, b, \delta), \\ y_j \left(\sum_{i=1}^n a_i x_{ij} + b \right) \geq 1 - \delta_j, \delta_j \geq 0, j = 1, \dots, N. \end{cases} \quad (17)$$

Этот критерий эквивалентен (13) при интерпретации независимых компонент случайного направляющего вектора $\mathbf{a} = (a_1, \dots, a_n)$ как распределенных по закону Лапласа $\psi(a_i | r) = (2r)^{-1/2} \exp(-(r/2)^{-1/2} |a_i|)$.

В диссертации не обсуждаются соответствующие алгоритмы оптимизации, они детально исследованы в работах^{13,14,16}. Заметим лишь, что в силу специфики принятых априорных распределений критерии (16) и (17) чрезвычайно важным свойством беспереборного отбора признаков непосредственно в ходе обучения. Это свойство является результатом склонности этих критериев в точке минимума присваивать нулевые значения большинству компонент направляющего вектора $a_i = 0$, реализуя при этом, вообще говоря, разумный отбор подмножества «полезных» признаков.

В диссертации теоретически исследована способность регуляризующих функций $\sum_{i=1}^n a_i^2$, $\sum_{i=1}^n a_i^2 + \mu \sum_{i=1}^n |a_i|$ и $\sum_{i=1}^n |a_i|$, соответственно, в классическом критерии SVM (5), Elastic Net SVM (16) и Lasso SVM (17), отбирать подмножество информативных признаков с точки зрения требований к селективности обучения, сформулированных на с. 4-5 автореферата. Исследование основано на сформулированном в диссертации понятии множеств наиболее и наименее предпочтительных ориентаций направляющего вектора разделяющей гиперплоскости в пространстве $\mathbb{R}^n$, определяющих «стиль» выбора подмножества полезных признаков.

Суждение о классическом SVM тривиально – все направления одинаковы по предпочтительности, и свойство селективности полностью отсутствует. Что же касается методов Elastic Net SVM и Lasso SVM, то оказалось, что они идентичны по ориентации предпочтительных направлений – наилучшим является направление, оставляющее единственный признак, а роль наихудшего играет направление, оценивающее все признаки как равнозначные. Это означает, что соответствующие априорные распределения направляющего вектора выражают предположение, что среди признаков есть только один признак, полезный для распознавания класса объекта. Следовательно, в ситуациях, когда в исходных данных таких признаков более одного, обучение методами Elastic Net SVM и Lasso SVM может приводить к снижению обобщающей способности полученных решающих функций.

¹⁵ Это утверждение доказано Е.О. Черноусовой.

¹⁶ Разин Н.А. Выпуклые критерии и параллелизуемые алгоритмы селективного комбинирования разнородных представлений объектов в задачах восстановления зависимостей по эмпирическим данным. Дисс. к.ф.-м.н. ВЦ РАН, 2013.

Известно также¹⁴, что Lasso SVM в группе линейно зависимых признаков полностью игнорирует факт их зависимости – отбирает лишь один такой признак, отбрасывая остальные. В то же время, Elastic Net SVM стремится оставлять всю линейно зависимую группу целиком, даже если она малоинформативна. В диссертации показано, что последнее свойство Elastic Net SVM также ведет к снижению обобщающей способности обучения.

Таким образом, необходима разработка новых способов регуляризации SVM, способных, во-первых, допускать существование независимых одинаково полезных признаков, и, во-вторых, удалять шумовые признаки без учета их линейной зависимости.

В третьей главе диссертации, в дополнение к существующим методам регуляризации, предложены два новых класса априорных распределений направляющего вектора, наделяющих обобщенный критерий обучения (13) свойством автоматического отбора признаков.

Метод релевантных признаков с регулируемой селективностью использует естественное обобщение классического метода опорных векторов за счет введения в его вероятностную постановку дополнительного предположения о том, что независимые компоненты направляющего вектора $\mathbf{a} = (a_1, \dots, a_n)$ распределены, как и прежде, по нормальному закону с нулевыми математическими ожиданиями, но с разными неизвестными случайными дисперсиями $r_i \geq 0, i = 1, \dots, n$, характеризующимися некоторым априорным распределением, и подлежащими оцениванию в процессе обучения по байесовскому принципу вместе с параметрами разделяющей гиперплоскости $(\mathbf{a}, b)$. В литературе подобные дополнительные априорные предположения о параметрах, участвующих в некотором ранее принятом априорном предположении, принято называть *Hyper Prior Assumptions*.

В этом случае совместная априорная плотность распределения направляющего вектора определяется выражением

$$\Psi(\mathbf{a} | \mathbf{r}) = \Psi(a_1, \dots, a_n | r_1, \dots, r_n) = \frac{1}{(2\pi)^{n/2} \prod_{i=1}^n r_i^{1/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n \frac{1}{r_i} a_i^2\right), \quad (18)$$

Удобно оперировать обратными дисперсиями $1/r_i$, называемыми мерами точности соответствующих случайных величин¹¹, полагая, что обратные дисперсии имеют одно и то же априорное гамма распределение $\gamma((1/r_i) | \alpha, \beta) \propto (1/r_i)^{\alpha-1} \exp(-\beta(1/r_i))$, $i = 1, \dots, n$. Будем выбирать значения двух параметров априорного гамма распределения α и β , задавая один общий параметр $0 < \mu < \infty$:

$$\alpha = (1 + \mu)^2 / 2\mu, \quad \beta = 1/2\mu.$$

В этом случае параметрическое семейство гамма распределений будет иметь вид

$$\gamma(1/r_i | \mu) \propto (1/r_i)^{(1+\mu^2)/2\mu} \exp\left[-(1/2\mu)(1/r_i)\right], \quad i = 1, \dots, n. \quad (19)$$

Нетрудно убедиться, что при увеличении параметра μ от 0 до ∞ вид априорного гамма распределения изменяется от сконцентрированного вокруг значения $1/r_i = \alpha/\beta = 1$ до почти «равномерного» на интервале $0 \leq 1/r_i < \infty$ (рис. 3.).

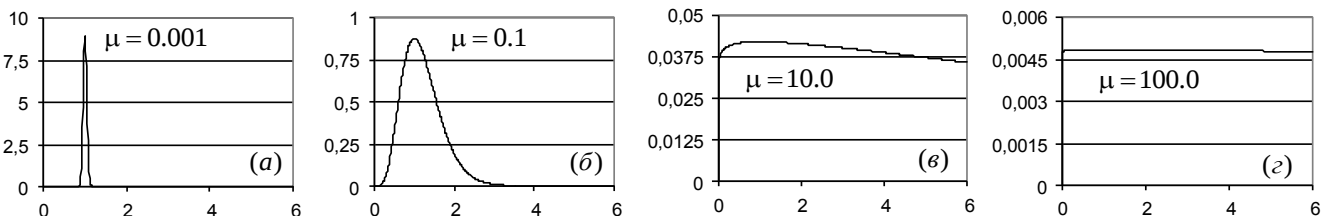


Рис. 3. Зависимость априорного гамма распределения дисперсий компонент направляющего вектора от параметра μ .

Обозначим через $G(\mathbf{r} | \mu) = G(r_1, \dots, r_n | \mu)$ априорную совместную плотность распределения обратных дисперсий:

$$G(\mathbf{r} | \mu) = G(r_1, \dots, r_n | \mu) = \prod_{i=1}^n \gamma((1/r_i) | \mu) \propto \left(\prod_{i=1}^n 1/r_i \right)^{(1+\mu^2)/2\mu} \exp\left(- (1/2\mu) \sum_{i=1}^n (1/r_i)\right). \quad (20)$$

Тогда совместная априорная плотность распределения параметров разделяющей гиперплоскости (18) и обратных дисперсий компонент направляющего вектора (20) запишется как

$$\bar{\Psi}(\mathbf{a}, b, \mathbf{r} | \mu) \propto \Psi(\mathbf{a} | \mathbf{r}) G(\mathbf{r} | \mu).$$

Соответственно, совместное апостериорное распределение всех переменных модели относительно обучающей совокупности, аналогично (11), пропорционально произведению

$$\bar{P}(\mathbf{a}, b, \mathbf{r} | X, Y; c, \mu) \propto \Psi(\mathbf{a} | \mathbf{r}) G(\mathbf{r} | \mu) \Phi(X | Y, \mathbf{a}, b, c).$$

Верхняя черта в обозначениях плотностей $\bar{\Psi}(\mathbf{a}, b, \mathbf{r} | \mu)$ и $\bar{P}(\mathbf{a}, b, \mathbf{r} | X, Y; c, \mu)$ отражает тот факт, что это совместные распределения (априорное и апостериорное) параметров гиперплоскости $(\mathbf{a}, b)$ и вектора дисперсий $\mathbf{r}$ в отличие от маргинальных распределений параметров гиперплоскости $\Psi(\mathbf{a}, b)$ (10) и $P(\mathbf{a}, b | X, Y; c)$ (11).

Отождествляя обучение, как и прежде, с вычислением байесовской оценки неизвестных параметров, мы придем к следующему критерию

$$(\hat{\mathbf{a}}, \hat{b}, \hat{\mathbf{r}} | X, Y; \mu, c) = \arg \max_{\mathbf{a} \in \mathbb{R}^n, b \in \mathbb{R}, \mathbf{r} \in \mathbb{R}^n} \bar{P}(\mathbf{a}, b, \mathbf{r} | X, Y, \mu, c) = \arg \max_{\mathbf{a} \in \mathbb{R}^n, b \in \mathbb{R}, \mathbf{r} \in \mathbb{R}^n} \left[\ln \Psi(\mathbf{a} | \mathbf{r}) + \ln G(\mathbf{r} | \mu) + \sum_{j=1}^N \ln \phi(\mathbf{x}_j | y_j; \mathbf{a}, b, c) \right], \quad (21)$$

представляющему собой обобщение критерия (12).

Теорема 3. Оптимизационная задача обучения (21) для (18), (19) и (20) эквивалентна критерию

$$\begin{cases} J_{RFM}(\mathbf{a}, b, \mathbf{r}, \delta | C, \mu) = \sum_{i=1}^n \left[(1/r_i) (a_i^2 + (1/\mu)) + ((1/\mu) + 1 + \mu) \ln r_i \right] + C \sum_{j=1}^N \delta_j \rightarrow \min(\mathbf{a}, b, \mathbf{r}, \delta), \\ y_j \left(\sum_{i=1}^n a_i x_{ij} + b \right) \geq 1 - \delta_j, \delta_j \geq 0, j = 1, \dots, N, \quad C = 2c. \end{cases} \quad (22)$$

При нулевом значении параметра $\mu \rightarrow 0$ и, соответственно, $1/\mu \rightarrow \infty$, штрафы $(1/\mu) [(1/r_i) + \ln r_i]$ настолько велики, что все оценки дисперсий равны единицы $r_1 = \dots = r_n = 1$ в точке минимума критерия (22), который становится эквивалентным критерию классического метода опорных векторов, вообще не обладающему свойством селективности.

При неограниченном увеличении параметра $\mu \rightarrow \infty$ априорное гамма распределение обратной дисперсии каждой компоненты направляющего вектора $1/r_i$ (19) и их совместное априорное распределение (20) становятся практически «равномерными», соответственно, на положительной полуоси $0 \leq 1/r_i < \infty$ (рис. 3-г) и в положительном квадранте пространства $\mathbb{R}^n$. Это в пределе приводит к исчезновению слагаемого $\ln G(\mathbf{r} | \mu)$ в критерии (21) и слагаемого $(1/\mu) [(1/r_i) + \ln r_i]$ в целевой функции критерия (22), штрафующего отклонение r_i от единицы, и критерий становится чрезмерно селективным.

В диссертации предложен алгоритм численного решения оптимизационной задачи (22), использующий метод Гаусса-Зайделя с поочередной оптимизацией по двум группам переменных $(a_1, \dots, a_n, b, \delta_1, \dots, \delta_N)$ и $(r_1, \dots, r_n)$. При фиксированных $(r_1, \dots, r_n)$, оптимизация по $(a_1, \dots, a_n, b, \delta_1, \dots, \delta_N)$ критерия (22) сводится к решению двойственной задачи квадратичного программирования

$$\begin{cases} W(\lambda | \mathbf{r}) = \sum_{j=1}^N \lambda_j - \sum_{j=1}^N \sum_{l=1}^n \lambda_j \lambda_l \left(\sum_{i=1}^n r_i x_{ij} x_{il} \right) y_j y_l \rightarrow \max(\lambda), \\ \sum_{j=1}^N y_j \lambda_j = 0, 0 \leq \lambda_j \leq C/2, j = 1, \dots, N. \end{cases} \quad (23)$$

Естественно принять начальные значения ($r_i^0 = 1, i = 1, \dots, n$), что соответствует стандартному методу опорных векторов (5). После того, как на очередном шаге итерационного процесса $k = 0, 1, 2, \dots$ найдено решение $(\lambda_1^k, \dots, \lambda_N^k)$ двойственной задачи (23) для текущего приближения $(r_1^k, \dots, r_n^k)$, новые значения дисперсий $(r_1^{k+1}, \dots, r_n^{k+1})$ на следующем шаге определяются следующей теоремой.

Теорема 4. Новые значения дисперсий $(r_1^{k+1}, \dots, r_n^{k+1})$ определяются решением $(\lambda_1^k, \dots, \lambda_N^k)$ двойственной задачи (23) для текущего приближения $(r_1^k, \dots, r_n^k)$ согласно выражению

$$r_i^{k+1} = \frac{(r_i^k)^2 \sum_{j: \hat{\lambda}_j^k > 0} \sum_{l: \hat{\lambda}_l^k > 0} y_j y_l x_{ij} x_{il} \hat{\lambda}_j^k \hat{\lambda}_l^k + 1/\mu}{\mu + 1 + 1/\mu}, \quad i = 1, \dots, n. \quad (24)$$

Итерационная процедура обучения обычно сходится за 10-15 шагов, а сам алгоритм демонстрирует способность подавлять неинформативные признаки за счет выбора очень маленьких весов $\hat{r}_i$ в решающем правиле

$$d(\mathbf{x} | \hat{\lambda}, \hat{b}, \hat{\mathbf{r}}) = \sum_{j: \hat{\lambda}_j > 0} y_j \hat{\lambda}_j \sum_{i=1}^n \hat{r}_i x_{ij} x_i + \hat{b} \geq 0,$$

где оптимальное значение порога $\hat{b}$ определяется выражением

$$\hat{b} = - \frac{\sum_{j: 0 < \hat{\lambda}_j < C/2} \hat{\lambda}_j \sum_{l: \hat{\lambda}_l > 0} y_l \hat{\lambda}_l \sum_{i=1}^n r_i x_{il} x_{ij} + (C/2) \sum_{j: \hat{\lambda}_j = C/2} y_j}{\sum_{j: 0 < \hat{\lambda}_j < C/2} \hat{\lambda}_j}.$$

Идея погружения дискретной задачи выбора подмножества в непрерывную задачу поиска в некотором смысле оптимальных неотрицательных весов, присваиваемых элементам исходного множества, заимствована из метода релевантных векторов Бишоп и Типпинга¹⁷, в котором направляющий вектор разделяющей гиперплоскости строится как взвешенная линейная комбинация векторов признаков всех объектов в обучающей совокупности. Векторы признаков объектов, получившие существенно ненулевые веса, названы Бишопом и Типпингом *релевантными векторами* в отличие от *опорных векторов* в методе опорных векторов, образующих жестко выделенное подмножество. В отличие от¹⁷, этот прием использован в диссертации для «мягкого» отбора признаков, а не объектов обучающей совокупности, участвующих в итоговом решающем правиле распознавания. Главным же отличием метода, предлагаемого в диссертации, является принципиально новое понятие параметра селективности μ .

В связи с этим обстоятельством предложенный в диссертации метод обучения (22) назван *методом релевантных признаков с регулируемой селективностью*, или, полностью характеризуя метод, *Selective Relevance Feature Support Vector Machine*.

Метод опорных признаков с регулируемой селективностью базируется на другом предположении об априорном распределении $\Psi(\mathbf{a})$ независимых компонент направляющего вектора $\mathbf{a} = (a_1, \dots, a_n)$. Вид априорной плотности $\Psi(\mathbf{a})$ представляет собой комбинацию распределения Лапласа при значениях нормы компонент, не превышающих заданного порога $|a_i| \leq \mu$, и нормального распределения при больших значениях $|a_i| > \mu$ (рис. 4):

$$\psi(a_i | \mu) \propto \exp(-q(a_i | \mu)), \quad q(a_i | \mu) = \begin{cases} 2\mu |a_i|, & |a_i| \leq \mu, \\ \mu^2 + a_i^2, & |a_i| > \mu. \end{cases} \quad (25)$$

¹⁷ C.M. Bishop, M.E. Tipping. Variational relevance vector machines. Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, 2000, pp. 46-53.

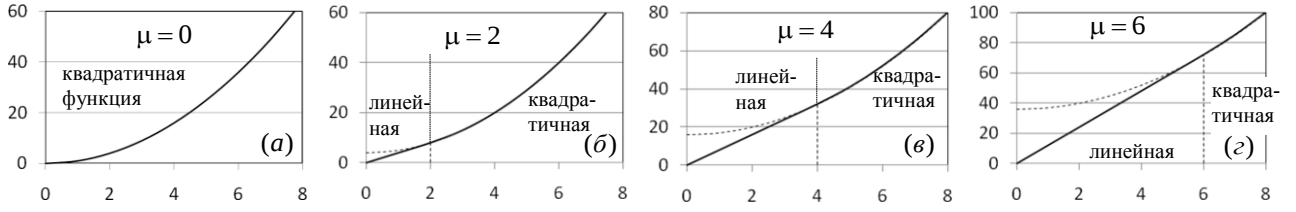


Рис. 4. Зависимость функции $q(a_i | \mu)$ комбинированного априорного распределения компонент направляющего вектора от параметра μ .

В этом случае совместное априорное распределение направляющего вектора $\mathbf{a} = (a_1, \dots, a_n)$ определяется выражением

$$\Psi(\mathbf{a}) = \Psi(a_1, \dots, a_n) \propto \exp\left(-\sum_{i=1}^n q(a_i | \mu)\right). \quad (26)$$

Априорное предположение (26) и обобщенный критерий обучения метода опорных векторов (13) определяют вид оптимизационной задачи обучения

$$\begin{cases} J_{SFM}(\mathbf{a}, b, \delta | C, \mu) = 2\mu \sum_{|a_i| \leq \mu} |a_i| + \sum_{|a_i| > \mu} (\mu^2 + a_i^2) + C \sum_{j=1}^N \delta_j \rightarrow \min(\mathbf{a}, b, \delta), \\ y_j \left(\sum_{i=1}^n a_i x_{ij} + b \right) \leq 1 - \delta_j, \delta_j \geq 0, j=1, \dots, N. \end{cases} \quad (27)$$

Задача (27) является задачей выпуклого программирования в пространстве $\mathbb{R}^n \times \mathbb{R}^{N+1}$, поскольку каждое слагаемое целевой функций выпукло в $\mathbb{R}$, а система ограничений для заданной обучающей выборки образует выпуклый многогранник.

Величина порога $0 \leq \mu < \infty$ плотности (25) выполняет роль параметра селективности обучения и при $\mu = 0 \Rightarrow q(a_i | \mu) = a_i^2$ оптимизационный критерий (27) эквивалентен критерию метода опорных векторов (5) с минимальной способностью к отбору признаков, а при достаточно больших значениях селективности $\mu \gg 0 \Rightarrow q(a_i | \mu) = 2\mu |a_i|$ - критерию Lasso SVM (17) с увеличивающейся селективностью обучения по мере увеличения μ относительно параметра c до полного подавления всех признаков.

Теорема 5. Решение задачи (27) определяется решением $(\hat{\xi}_i \geq 0, i \in I = \{1, \dots, n\}, \hat{\lambda}_j \geq 0, j=1, \dots, N)$ двойственной задачи

$$\begin{cases} L(\lambda_1, \dots, \lambda_N | C, \mu) = \sum_{j=1}^N \lambda_j - \sum_{i \in I} \sum_{l=1}^n (1/2) \xi_i \rightarrow \max(\lambda_1, \dots, \lambda_N), \\ \xi_i \geq 0, \xi_i \geq \sum_{j=1}^N \sum_{l=1}^n y_j y_l x_{ij} x_{il} \lambda_j \lambda_l - \mu^2, i \in I = \{1, \dots, n\}, \\ \sum_{j=1}^N y_j \lambda_j = 0, 0 \leq \lambda_j \leq C/2, j=1, \dots, N. \end{cases} \quad (28)$$

и выражается следующим образом

$$\begin{cases} \hat{a}_i = 0, & i \in I^- = \left\{ i \in I : \sum_{j=1}^N \sum_{l=1}^n y_j y_l x_{ij} x_{il} \hat{\lambda}_j \hat{\lambda}_l - \mu^2 < 0 \right\}, \\ \hat{a}_i = \hat{\eta}_i \sum_{j: \hat{\lambda}_j > 0} y_j \hat{\lambda}_j x_{ij}, & i \in I^0 = \left\{ i \in I : \sum_{j=1}^N \sum_{l=1}^n y_j y_l x_{ij} x_{il} \hat{\lambda}_j \hat{\lambda}_l - \mu^2 = 0 \right\}, \\ \hat{a}_i = \sum_{j: \hat{\lambda}_j > 0} y_j \hat{\lambda}_j x_{ij}, & i \in I^+ = \left\{ i \in I : \sum_{j=1}^N \sum_{l=1}^n y_j y_l x_{ij} x_{il} \hat{\lambda}_j \hat{\lambda}_l - \mu^2 > 0 \right\}. \end{cases} \quad (29)$$

Теорема 6. Итоговое решение о классе нового объекта имеет вид

$$d(\mathbf{x} | \hat{\lambda}, \eta_i, i \in I^0, b) = \sum_{j: \hat{\lambda}_j > 0} y_j \hat{\lambda}_j \left[\sum_{i \in I^+} x_{ij} x_i + \sum_{i \in I^0} \hat{\eta}_i x_{ij} x_i \right] + \hat{b} \geq 0, \quad (30)$$

где коэффициенты $\hat{\eta}_i$ определяются дополнительной задачей линейного программирования

$$\begin{cases} 2\mu \sum_{i \in I^0} \eta_i + C \sum_{j=1}^N \delta_j \rightarrow \min(\eta_i, i \in I^0, b, \delta | \lambda), \\ \sum_{i \in I^+} \left[\sum_{l=1}^N y_j y_l x_{ij} x_{il} \hat{\lambda}_l \right] \eta_i + y_j b + \delta_j \geq 1 - \sum_{i \in I^+} \sum_{l=1}^N y_j y_l x_{ij} x_{il} \hat{\lambda}_l, \\ \delta_j \geq 0, \quad j=1, \dots, N, \quad 0 \leq \eta_i \leq 1, \quad i \in I^0. \end{cases} \quad (31)$$

Оптимальные ненулевые значения компонент $a_i \in \mathbb{R}$, $i \in I^0 \cup I^+$ направляющего вектора (29), как и в классическом методе опорных векторов, выражаются в виде линейной комбинации опорных объектов обучающей выборки (2), но, в отличие от SVM, решение задачи (29) явным образом указывает множество $i \in I^-$ строго нулевых компонент $a_i = 0$, а, следовательно, множество неактивных или неопорных признаков, т.е. признаков, не участвующих в принятии решения (30) о классе нового объекта. По этой причине, по аналогии с методом опорных векторов, предложенный метод назван в диссертации *методом опорных признаков с регулируемой селективностью* или *Selective Support Feature Support Vector Machine*.

Параметр селективности $\mu \geq 0$, как и параметр $C \geq 0$, является структурным параметром критериев обучения (22) и (27), и определяет последовательность вложенных классов моделей распознавания, «размерность» которых уменьшается с ростом μ . Для выбора оптимальных значений параметров в данной работе используется метод скользящего контроля, зарекомендовавший себя, как наиболее универсальный и один из самых эффективных способов оценивания обобщающей способности методов обучения по обучающей выборке.

Преимущества предложенных двух критериев селективного обучения метода релевантных признаков (22) и метода опорных признаков (27) по сравнению с известными критериями Elastic Net SVM (16) и Lasso SVM (17) исследуются в диссертации с теоретической точки зрения в конце главы 3 и экспериментально в главе 4.

В частности, теоретическое исследование регуляризующей способности метода опорных признаков (27) показало, что множество наименее предпочтительных ориентаций направляющего вектора состоит не из одного направления, оценивающего все признаки как равнозначные (методы Elastic Net SVM и Lasso SVM), а составляет конус направлений

$\{\mathbf{a} \in \mathbb{R}^n : \sum_{i=1}^n a_i^2 = 1, |a_i| \geq \varepsilon \mu, i=1, \dots, n\}$, предполагающих включение в решающую функцию $\mathbf{a}^T \mathbf{x} + b \geq 0$ сразу всех признаков с разными достаточно большими весами $|a_i| \geq \varepsilon \mu$, $i=1, \dots, n$.

Это условие означает «неразличимость» для обучения таких направляющих векторов между собой с точки зрения их априорной предпочтительности. Следовательно, в ситуациях, когда в исходных данных есть несколько значимых признаков, выбор весов, с которыми они будут входить в решение, определяется исключительно качеством разделения выборки, а не априорным стремлением выделить только один из признаков, подобно методам Elastic Net SVM и Lasso SVM.

С другой стороны, предложенный метод опорных признаков (27) характеризуется избирательным «стилем» учета линейной зависимости признаковых описаний объектов при их отборе. Избирательность предложенного метода выражается в стремлении сохранять в итоговом решении только попарно зависимые признаки, обладающие достаточной значимостью для распознавания. Для сравнения метод Elastic Net SVM сохраняет все линейно зависимые признаки. Малозначимые признаки при этом максимально подавляются в процессе обучения независимо от их попарной линейной зависимости, аналогично Lasso SVM.

В четвертой главе приводятся результаты экспериментального исследования предложенных в диссертации методов обучения с регулируемой селективностью, а именно, метода релевантных признаков (22) и метода опорных признаков (27). Основной целью экспериментального исследования является анализ предложенных методов в сравнении с существующими методами SVM (5), Lasso SVM (17) и Elastic Net SVM (16) по их способности сокращать признаковое описание объектов распознавания и, в конечном итоге, повышать

обобщающую способность обучения при относительно малой обучающей выборке и большом числе признаков.

Экспериментальное исследование имеет общепринятую структуру, включающую в себя серию модельных экспериментов и пример решения прикладной задачи. Модельные эксперименты диссертации основаны, по своей структуре, на исследовании, проведенном авторами метода Elastic Net SVM¹⁴ для иллюстрации преимущества модульно-квадратичной функции штрафа Elastic Net по сравнению с традиционным модульным Lasso. Прикладной задачей является задача распознавания рака легких из репозитория UCI¹⁸.

С одной стороны, преимущество модельных экспериментов заключается в том, что они дают возможность придать условным понятиям полезных и лишних признаков абсолютный смысл, и позволяют для каждой конкретной гиперплоскости непосредственно вычислить вероятность ошибки распознавания на генеральной совокупности. С другой стороны, это позволяет сравнивать методы в заведомо разных условиях порождения исходных данных, скрытых для наблюдения в прикладных задачах и выраженных в знании истинного положения искомой гиперплоскости и особенностях распределений признакового описания объектов.

Именно этот аспект организации экспериментального исследования заслуживает особого внимания. Структура четырех модельных задач настоящей диссертации соответствует четырем простым требованиям к селективному обучению, приведенным на с. 5 автореферата. Результаты модельных экспериментов, изложенные в диссертации, наглядно иллюстрируют весьма важный факт, что ни один из существующих методов селективного обучения, включая предлагаемые, не удовлетворяет сразу всем этим весьма неизощренным требованиям. Предложенный метод опорных признаков (27) хорошо справился с подавлением шумовых признаков (а,б) при линейно независимых информативных признаках (в), существенно улучшив и без того неплохой результат существующего метода Elastic Net SVM. Однако метод релевантных признаков (22) в этих условиях оказался далеко не так эффективен. Требование выделять группу информативных признаков, которые только вместе обеспечивают достаточную точность распознавания (г), оказалось весьма проблематичным, как для существующего Elastic Net SVM, так и для предложенного метода опорных признаков, селективность которых обеспечивается за счет использования в целевых функциях критериев обучения штрафа модуля. Вместе с тем требование (г) не доставило никаких сложностей для второго предложенного метода релевантных признаков, селективность которого имеет отличную от привычного модуля природу.

Полученные в рамках диссертационной работы результаты экспериментального исследования иллюстрируют полезность предложенных методов обучения в сравнении с известными образцами.

В заключении приведены основные результаты диссертационной работы. **В приложениях** приведены доказательства теорем, сформулированных в диссертации.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ДИССЕРТАЦИИ

1. Сформулирована общая математическая постановка задачи обучения двухклассовому распознаванию образов в линейном пространстве признаков на основе количественного измерения расстояния между вектором признаков объекта и разделяющей гиперплоскостью.

2. Сформулирована байесовская постановка задачи обучения двухклассовому распознаванию образов в линейном пространстве признаков, дающая вероятностное обоснование метода опорных векторов.

3. Сформулированы две новые частные априорные вероятностные модели генеральной совокупности в терминах совместного априорного распределения компонент направляющего вектора разделяющей гиперплоскости.

¹⁸ <http://archive.ics.uci.edu/ml/datasets/Lung+Cancer> – UCI Machine Learning Repository: Lung Cancer Data Set.

4. Разработаны два метода и два соответствующих алгоритма апостериорного оценивания разделяющей гиперплоскости, реализующих байесовский принцип обучения с заданной селективностью отбора признаков при выбранных априорных предположениях о генеральной совокупности объектов двух классов. Разработанные методы получили названия метод релевантных признаков и метод опорных признаков с регулируемой селективностью.

5. Теоретически исследованы преимущества предложенных двух критериев селективного обучения метода релевантных признаков и метода опорных признаков по сравнению с известными критериями Elastic Net SVM и Lasso SVM.

6. Экспериментально получены количественные оценки эффективности предложенных методов обучения на модельных данных и данных задачи распознавания рака легких.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Середин О.С., Моттль В.В., Татарчук А.И., Разин Н.А. Выпуклые селективные критерии метода релевантных векторов в пространстве парных отношений объектов распознавания. Известия ТулГУ, Серия Естественные науки. Тула: Изд-во ТулГУ, 2013, Вып. 1, с. 165-176.

2. O. Seredin, V. Mottl, A. Tatarchuk, N. Razin, D. Windridge. Convex Support and Relevance Vector Machines for selective multimodal pattern recognition. Proceedings of the 21th International Conference on Pattern Recognition, Tsukuba, Japan, November 11-15, 2012. IAPR, 2012, ISSN 978-4-9906441-1-6, 2012, pp. 1647-1650.

3. A. Tatarchuk, E. Urlov, V. Mottl, D. Windridge. A support kernel machine for supervised selective combining of diverse pattern-recognition modalities. In: Multiple Classifier Systems. Lecture Notes In Computer Science, Vol. 5997. Springer-Verlag, Berlin \ Heidelberg, 2010, ISBN 978-3-642-12127-2_17, pp. 165-174.

4. Татарчук А.И., Урлов Е.Н., Моттль В.В. Метод опорных потенциальных функций в задаче селективного комбинирования разнородной информации при обучении распознаванию образов. Доклады 14-й Всероссийской конференции «Математические методы распознавания образов», Суздаль, 21-26 сентября 2009 г., с. 192-195.

5. Татарчук А.И., Урлов Е.Н., Ляшко А.С., Моттль В.В. Экспериментальное исследование обобщающей способности методов селективного комбинирования потенциальных функций в задаче двухклассового распознавания образов. Доклады 14-й Всероссийской конференции «Математические методы распознавания образов», Суздаль, 21-26 сентября 2009 г., с. 196-199.

6. Татарчук А.И., Сулимова В.В., Моттль В.В., Уиндридж Д. Метод релевантных потенциальных функций для селективного комбинирования разнородной информации при обучении распознаванию образов на основе байесовского подхода. Доклады 14-й Всероссийской конференции «Математические методы распознавания образов», Суздаль, 21-26 сентября 2009 г., с. 188-191.

7. A. Tatarchuk, V. Sulimova, D. Windridge, V. Mottl, M. Lange. Supervised selective combining pattern recognition modalities and its application to signature verification by fusing on-line and off-line kernels. In: Multiple Classifier Systems. Lecture Notes In Computer Science, Vol. 5519. Springer-Verlag, Berlin \ Heidelberg. ISBN 978-3-642-02325-5, 2009, pp. 324-334.

8. A. Tatarchuk, V. Mottl, A. Elisseyev, D. Windridge. A Bayesian approach to multimodal pattern recognition with supervised selectivity of modalities. Proceedings of 9th International Conference on Pattern Recognition and Image Analysis: New Information Technologies, Nizhni Novgorod, September 14-20, 2008, vol. 2, p. 204.

9. В.В. Моттль, А.И. Татарчук, А.П. Елисеев Экспериментальное исследование методов многомодального распознавания образов с регулируемой селективностью.

Известия Тульского государственного университета. Технические науки, вып.3. – Тула: Изд-во ТулГУ, 2008.

10. Tatarchuk A., Mottl V., Elisyev A., Windridge D. Selectivity supervision in combining pattern-recognition modalities by feature- and kernel-selective Support Vector Machines. Proceedings of the 19th International Conference on Pattern Recognition, Vol 1-6, IEEE, ISBN 978-1-4244-2174-9, 2008, pp. 2336-2339.

11. Моттль В.В., Татарчук А. И., Елисеев А. П. Регулируемая селективность в многомодальном распознавании образов. Таврический вестник информатики и математики, 2008, том. 2, с. 89.

12. Моттль В.В., Татарчук А. И., Елисеев А. П. Многомодальное распознавание образов с регулируемой селективностью. Тезисы докладов международной научной конференции «Интеллектуализации обработки информации», Симферополь, 2008, с. 172.

13. Моттль В.В., Татарчук А. И., Елисеев А. П. Исследование методов комбинирования классификаторов и потенциальных функций в многомодальном распознавании образов. Тезисы докладов международной научной конференции «Интеллектуализации обработки информации», Симферополь, 2008, с. 176.

14. Mottl V., Tatarchuk A., Krasotkina O., Seredin O., Sulimova V. Combining pattern recognition modalities at the sensor level via kernel fusion. In: Multiple Classifier Systems. Lecture Notes In Computer Science, Vol. 4472. Springer-Verlag, Berlin \ Heidelberg. ISBN 978-3-540-72481-0, 2007, pp. 1-12.

15. Моттль В.В., Татарчук А. И., Красоткина О.В., Сулимова В.В. Комбинирование потенциальных функций в многомодальном распознавании образов. Математические методы распознавания образов. 13-я Всероссийская конференция: Сборник докладов. М.: МАКС Пресс, 2007.

16. Valentina Sulimova, Vadim Mottl, Alexander Tatarchuk Multi-Kernel approach to on-line signature verification. // Proceedings of the Eighth IASTED International Conference on Signal and Image Processing, ACTA PRESS ANAHEIM, ISBN 978-0-88986-583-9, 2006, pp. 448-453.

17. В.В. Сулимова, В.В. Моттль, А.И. Татарчук. Автоматический выбор наиболее информативных фрагментов в задачах распознавания сигналов разной длительности. // Труды конференции «Интеллектуализация обработки информации»-2006, Алушта, стр. 152-154.

18. В.В. Сулимова, В.В. Моттль, А.И. Татарчук. Автоматический выбор наиболее информативных фрагментов в задачах распознавания сигналов разной длительности. // Таврический вестник математики и информатики – № 1, 2006, стр. 109-115.

Жирным шрифтом выделены статьи в изданиях, рекомендованные ВАК.

Личный вклад соискателя в работах с соавторами заключается в следующем:

[1,2] Предложен выпуклый селективный критерий релевантных векторов и исследованы свойства двойственной задачи обучения.

[3,4] Предложен критерий метода опорных признаков и опорных потенциальных функций, исследованы свойства двойственной задачи обучения, разработан алгоритм решения выпуклой задачи обучения.

[5,9,13] Разработана структура экспериментального исследования методов релевантных признаков и опорных признаков.

[6] Разработан байесовский подход к задаче обучения распознаванию образов по методу опорных векторов с обобщением на метод релевантных признаков и потенциальных функций.

[7,8] Разработана вероятностная методология селективного комбинирования разнородных модальностей представления объектов распознавания.

[10,11] Разработан механизм регулирования селективности отбора признаков при обучении распознаванию образов по методу опорных векторов.

[14,15] Разработаны принципы комбинирования разнородных представлений объектов на уровне сенсоров.

[16] Разработана математическая идея применения методов отбора признаков для задачи распознавания рукописных подписей, представленных динамическими характеристиками.

[17,18] Разработана математическая идея применения методов отбора признаков для выделения информативных фрагментов сигналов.

A handwritten signature in black ink, appearing to read 'Машарук'.