На правах рукописи

Разин Николай Алексеевич

ВЫПУКЛЫЕ КРИТЕРИИ И ПАРАЛЛЕЛИЗУЕМЫЕ АЛГОРИТМЫ СЕЛЕКТИВНОГО КОМБИНИРОВАНИЯ РАЗНОРОДНЫХ ПРЕДСТАВЛЕНИЙ ОБЪЕКТОВ В ЗАДАЧАХ ВОССТАНОВЛЕНИЯ ЗАВИСИМОСТЕЙ ПО ЭМПИРИЧЕСКИМ ДАННЫМ

Специальность 05.13.17 — Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание учёной степени

кандидата физико-математических наук

Москва — 2013

Работа выполнена на кафедре «Интеллектуальные системы» факультета управления и прикладной математики Федерального государственного автономного образовательного учреждения высшего профессионального образования «Московский физико-технический институт (государственный университет)».

Научный руководитель: доктор технических наук, профессор **Моттль Вадим Вячеславович**.

Официальные оппоненты: доктор физико-математических наук, доцент **Устинин Михаил Николаевич**.
Место работы: Федеральное государственное бюджетное учреждение науки Институт математических проблем биологии РАН, заместитель директора по научной работе.

кандидат технических наук, доцент **Копылов Андрей Валериевич**.
Место работы: Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования Тульский государственный университет.

Ведущая организация: **Федеральное государственное бюджетное учреждение науки Институт проблем управления РАН.**

Защита состоится 19 декабря 2013 г. в 15-30 на заседании диссертационного совета Д 002.017.02 при Федеральном государственном бюджетном учреждении науки «Вычислительный центр им. А.А. Дородницына Российской академии наук», расположенном по адресу: 119333, г.Москва, улица Вавилова, 40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан 11 ноября 2013 года.

Ученый секретарь диссертационного совета
Д 002.017.02, д.ф.-м.н., профессор  Рязанов В.В.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Пожалуй, самой популярной задачей современной информатики является задача восстановления объективно существующей зависимости между наблюдаемыми свойствами определённого вида объектов реального мира и их некоторой их скрытой характеристикой, доступной для наблюдения лишь в пределах конечной обучающей совокупности. Такую задачу принято называть задачей обучения. Простейшим и наиболее изученным видом представления объектов в задачах восстановления зависимостей по эмпирическим данным является их представление в виде совокупности действительных признаков, что позволяет использовать хорошо разработанные линейные методы анализа данных. Каждый признак выражает определённый вид информации об объектах реального мира, доступной наблюдателю, который принято называть модальностью представления объектов. Естественное стремление не пропустить свойства объекта, фактически важные для предсказания его целевой скрытой характеристики, заставляет наблюдателя выражать числовыми признаками как можно большее число модальностей представления объектов. Однако, если число признаков приближается к числу объектов в обучающей совокупности, то резко ухудшается обобщающая способность обучения, т.е. возможность применения модели зависимости, полученной в результате обучения, к остальным объектам, не участвовавшим в обучении. Чтобы избежать этого эффекта, называемого переобучением, необходимо выбрать лишь некоторое подмножество наиболее информативных признаков и отбросить остальные. Эта задача, называемая задачей отбора полезных признаков, остаётся наиболее проблематичной задачей в современной теории обучения. Во многих практических ситуациях не удаётся указать числовые признаки объектов, адекватно выражающие их свойства, связанные со значениями ненаблюдаемой целевой характеристики. Вместо поиска индивидуальных свойств объектов, предположительно полезных для предсказания значения их целевой характеристики, природа изучаемого явления часто подсказывает способ их попарного сравнения, например, в виде степени похожести или непохожести в некотором смысле, так что похожие объекты, скорее всего, будут иметь более близкие значения целевой характеристики, чем непохожие. Тогда, если сравнение двух объектов выражается действительным числом, то естественным способом представления отдельного объекта является совокупность результатов его сравнения со всеми объектами обучающей совокупности. Получающийся числовой вектор принято называть вектором вторичных признаков объекта. Поскольку в этом случае число признаков совпадает с размером обучающей совокупности, то неизбежно возникает проблема отбора подмножества так называемых релевантных объектов, с которыми достаточно сравнивать каждый новый объект. Понятно, что такая задача практически совпадает с задачей отбора подмножества полезных индивидуальных признаков объектов. Здесь роль модальности представления объектов фактически играет выбранный способ их парного сравнения. Типичны ситуации, когда существуют несколько аль-

тернативных способов парного сравнения объектов (модальностей сравнения). Тогда в качестве вектора вторичных признаков объекта естественно рассматривать результаты его сравнения со всеми объектами обучающей совокупности сразу по всем модальностям. Естественно, что при этом проблема отбора вторичных признаков многократно усложняется - во-первых, их число в несколько раз превышает объём обучающей совокупности, и, во-вторых, теперь отбирать придется не только релевантные объекты, но и релевантные модальности сравнения для каждого из них. В данной работе исследуются две принципиальных проблемы, возникающих при попытке отбирать релевантные объекты и релевантные модальности.

Первая проблемная ситуация заключается в том, что большинство известных способов отбора полезных признаков являются дискретными и приводят к задачам комбинаторной оптимизации, обладающим слишком высокой вычислительной сложностью. Попытки упрощения алгоритмов переборного выбора подмножества признаков, как правило, приводят к снижению качества классификации на генеральной совокупности. С другой стороны, известные методы замены исходной задачи дискретной оптимизации подходящей непрерывной задачей приводят к невыпуклым критериям обучения.

Для разрешения этой проблемной ситуации в работе предлагается класс непрерывных выпуклых критериев обучения, специально сконструированных для задач восстановления зависимостей на основе нескольких конкурирующих функций парного сравнения. Предлагаемые выпуклые критерии определяют для всякой обучающей совокупности единственное подмножество релевантных объектов для сравнительного представления всякого нового объекта, а также подмножество релевантных функций сравнения.

Вторая проблемная ситуация заключается в том, что при больших размерах обучающей совокупности и при большом числе пробных модальностей парного сравнения объектов численное решение даже выпуклой задачи обучения на однопроцессорных вычислительных машинах занимает слишком много времени. Возникает необходимость использования параллельных вычислительных систем, что требует специального построения процедуры обучения.

В основе предлагаемого пути преодоления второй проблемной ситуации лежит идея разработки хорошо масштабирующихся по мощности вычислительных систем алгоритмов. Алгоритмы решения неквадратичных выпуклых задач оптимизации, типичных для восстановления зависимостей по эмпирическим данным, неизбежно являются итерационными. Существующие алгоритмы в стандартных случаях делают большое количество «быстрых» итераций, поэтому увеличение количества процессоров не приводит к увеличению в такое же количество раз скорости работы алгоритма. В работе предлагается использовать алгоритмы, делающие небольшое количество итераций, каждую из которых можно эффективно распараллелить. Таким образом, предлагаемые в работе алгоритмы будут хорошо масштабироваться на мощных многопроцессорных вычислительных системах.

Объект исследования: прикладные задачи распознавания образов, в которых единственным способом восприятия объекта распознавания является измерение его сходства/несходства с другими объектами.

Предмет исследования: методы обучения распознаванию образов и восстановления числовых зависимостей на базе измерения сходства/несходства между объектами обучающей совокупности.

Цель диссертации. Целью диссертации является разработка алгоритмически эффективных методов выбора подмножества базисных объектов и подмножества способов сравнения в задачах распознавания образов и восстановления числовых зависимостей.

Задачи исследования. Для реализации изложенных основных принципов предлагаемого подхода к построению выпуклых критериев и параллелизуемых алгоритмов селективного комбинирования разнородных представлений объектов в задачах восстановления зависимостей по эмпирическим данным в данной работе ставятся следующие основные задачи:

1. Разработка выпуклых критериев отбора релевантных подмножеств объектов обучающей совокупности и функций парного сравнения с регулируемой селективностью в задачах обучения распознаванию образов и восстановления числовых зависимостей.

2. Разработка итерационных алгоритмов решения задач обучения распознаванию образов и восстановления числовых зависимостей с заданной селективностью отбора релевантных объектов и релевантных функций парного сравнения.

3. Разработка многопроцессорных алгоритмов реализации отдельной итерации обучения распознаванию образов и восстановления числовых зависимостей с заданной селективностью

4. Разработка последовательных алгоритмов выбора уровня селективности в задачах обучения на основе парного сравнения объектов.

5. Создание программно-алгоритмического комплекса, реализующего алгоритмы селективного комбинирования разнородных представлений объектов, хорошо масштабирующихся на большое число процессоров.

6. Экспериментальное исследование разработанных алгоритмов обучения распознаванию образов и восстановления числовых зависимостей на основе селективного комбинирования функций парного сравнения объектов.

Методы исследования. Исследование базируется на использовании классической теории распознавания образов, методов оптимизации, теории линейных пространств со скалярным произведением.

Научная новизна. В работе впервые предложена концепция отбора релевантных объектов и релевантных способов сравнения объектов одновременно.

Предложен класс выпуклых критериев обучения беспризнаковому распознаванию образов и восстановления числовых зависимостей, решающее правило которых содержит небольшую долю релевантных объектов и релевантных способов парного сравнения. Предложен эффективный алгоритм регуляризации для задач обучения беспризнаковому распознаванию образов. Предложен эффективный, хорошо масштабирующийся на многопроцессорные системы алгоритм решения выпуклых задач оптимизации, рассматриваемых в данной работе. Приведена реализация алгоритма параллельного умножения матриц на видеокарте, ускоряющая наивный алгоритм умножения матриц в 25 раз.

Положения, выносимые на защиту.

1. Концепция отбора релевантных подмножеств объектов обучающей совокупности и функций парного сравнения с регулируемой селективностью в задачах обучения распознаванию образов и восстановления числовых зависимостей.

2. Выпуклые критерии, отбирающие релевантные объекты и релевантные функции парного сравнения в задачах обучения распознаванию образов и восстановления числовых зависимостей.

3. Эффективные последовательные алгоритмы выбора уровня селективности в задачах обучения на основе парного сравнения объектов.

4. Эффективные последовательные алгоритмы в задачах восстановления числовых зависимостей на основе парного сравнения объектов.

5. Эффективные, хорошо масштабирующиеся на многопроцессорные системы алгоритмы реализации отдельной итерации обучения распознаванию образов и восстановления числовых зависимостей с заданной селективностью.

Достоверность полученных результатов подтверждается доказательствами сформулированных в диссертации теорем и результатами решения прикладных задач.

Практическая значимость. Разработанные концепции и методы позволяют строить решающие правила распознавания образов в множествах объектов, для которых проблематично формирование индивидуальных векторов признаков. При этом сами решающие правила включают в себя сравнительно небольшое количество релевантных объектов и релевантных функций парного сравнения, что значительно повышает их интерпретируемость и делает эффективным определение целевой характеристики для нового произвольного объекта. Предложенные эффективные алгоритмы позволяют применять разработанные методы решения задач распознавания образов и восстановления числовых зависимостей на многопроцессорных системах, значительно расширяя спектр задач, прежде всего позволяя обрабатывать большие объёмы данных.

Связь с плановыми научными исследованиями. Работа выполнена при поддержке грантов Российского фонда фундаментальных исследований №№ 11-07-00409-а, 11-07-00634-а, 12-07-13142-офи-м.

Реализация и внедрение результатов работы. Результаты исследования применены для решения задачи распознавания вторичной структуры белков по составляющим их последовательностям аминокислот, и задачи восстановления числовой зависимости намагниченности полупроводника от времени эксперимента.

Апробация работы. Основные положения и результаты диссертации докладывались на конференциях: «Интеллектуализация обработки информации» (Будва, Черногория, 2012 г.), «Распознавание образов в биоинформатике» (PRIB-2012, Токио, Япония, 2012 г.), «Международная конференция по распознаванию образов» (ICPR-2012, Цукуба, Япония, 2012 г.), «Математические методы распознавания образов» (Казань, 2013 г.).

Публикации. По тематике работы опубликовано 7 статей, в том числе 2 статьи в журналах, рекомендованных ВАК.

Структура и объём работы. Диссертация состоит из введения, пяти глав, заключения, списка литературы. Материал изложен на 105 страницах, содержит 9 рисунков, 4 таблицы, список литературы из 43 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность исследований по разработке методов и алгоритмов беспризнакового распознавания образов, цели и задачи проводимых исследований, положения, выносимые на защиту, приведены сведения о структуре диссертации, её апробации и внедрении.

В первой главе проанализирован круг прикладных задач, приводящих к необходимости обучения при отсутствии явного вектора признаков и необходимости отбора релевантных объектов и релевантных методов парного сравнения объектов. Проанализированы классические методы отбора признаков в задачах восстановления зависимостей по эмпирическим данным. Проанализированы классические методы беспризнакового восстановления зависимостей по эмпирическим данным. Сформулированы основные проблемы существующих методов восстановления зависимостей в случае отсутствия явного вектора признаков.

Алгоритм вычисления оценок, предложенный Ю.И. Журавлёвым еще в 70-х годах, как раз ставит задачу отбора подмножества реле-вантных объектов (опорного подмножества объектов), описывая каждый объект множеством значений разных функций парного сравнения. Но, в силу недостатка вычислительных мощностей, для решения этой задачи используется комбинаторный подход, а не оценивается разделяющая гиперплоскость. Впоследствии Роберт Дьюин предложил использовать не сами признаки в качестве описания объектов, а значения некоторой функции парного сравнения $S(\omega', \omega'')$ данного объекта со всеми объектами обучающей совокупности $\omega_j \in \Omega = \omega_j, j = \overline{1, N}$.

В данной работе рассматривается ситуация, когда функций парного сравнения, описывающих объекты, задано несколько: $S_i(\omega', \omega''), i = \overline{1,m}$. Из этих функций нужно выбрать подмножество релевантных. Соответственно, каждый объект теперь описывается mN числами, то есть мы имеем ситуацию, когда количество признаков, которыми характеризуется объект, значительно превышает объём обучающей совокупности. Поэтому очень важно отобрать из этого множества признаков релевантные, чтобы избежать эффекта переобучения и чтобы процесс распознавания нового объекта не требовал большого количества памяти и вычислительных ресурсов.

Начнём с задачи обучения распознаванию двух классов объектов. Для того, чтобы эффективно отбирать из заданного множества вторичных признаков релевантные, предлагается использовать критерий SVM с квадратично-модульной регуляризацией. Как и в классической SVM, предполагается, что объекты реального мира $\omega \in \Omega$ описываются n признаками $\mathbf{x}(\omega) = (x_1(\omega), ..., x_n(\omega))$, которых значительно больше, чем объектов N в обучающей выборке, в нашем случае $n = mN$. Исследуется следующая задача поиска оптимальной разделяющей гиперплоскости $y(\mathbf{x}, \mathbf{a}) = \sum_{i=1}^{n} a_i x_i + b \lesseqgtr 0$ для заданной обучающей совокупности:

$$\begin{cases} \sum_{i=1}^{n} [\beta a_i^2 + \mu |a_i|] + \sum_{j=1}^{N} \delta_j \to \min\limits_{\mathbf{a}, b, \delta} \\ y_j (\sum_{i=1}^{n} a_i x_{ji} + b) \geqslant 1 - \delta_j, j = \overline{1,N} \\ \delta_j \geqslant 0, j = \overline{1,N} \end{cases} \tag{1}$$

Коэффициенты β, μ должны удовлетворять условиям: $\beta \geqslant 0, \mu \geqslant 0, \beta + \mu > 0$. Критерий остаётся строго выпуклым, гарантируя единственность решения.

Критерий обучения (1) остаётся полностью применимым в ситуации, когда представление объектов при помощи функцию сравнения $S(\omega', \omega'')$ более удобно, чем при помощи признакового описания $\mathbf{x}(\omega)$. Значения функции сравнения объекта $\omega \in \Omega$ на каждом из N объектов обучающей совокупности $x_l(\omega) = S(\omega_l, \omega)$ могут быть использованы как его вторичные признаки. В этом случае мы получаем обобщённую версию RVM – Relevance Vector Machine, впервые предложенную Кристофером Бишопом и Майклом Типпингом, которую в нашей версии следует назвать Relevance Object Machine, потому что нет никакой связи с представлением объектов при помощи векторов признаков.

Возможность представления объектов несколькими априори равновероятными функциями парного сравнения $S_i(\omega', \omega''), i = \overline{1,m}$ не влияет принципиально на критерий, кроме того, что количество вторичных признаков расширяется до mN:

$$\begin{cases} x_{il} = S_i(\omega_l, \omega) \text{для всех } \omega \in \Omega \\ x_{il,j} = S_i(\omega_l, \omega_j) \text{для } j\text{-ого объекта } \omega_j \end{cases} \tag{2}$$

Прямое обобщение критерия (1) даёт выпуклый критерий обучения, который отличается только количеством переменных:

$$\begin{cases} \sum\limits_{i=1}^{m}\sum\limits_{l=1}^{N}\left[\beta a_{il}^2 + \mu|a_{il}|\right] + \sum\limits_{j=1}^{N}\delta_j \to \min\limits_{a_{il},b,\delta_j}, \\ y_j(\sum\limits_{i=1}^{m}\sum\limits_{l=1}^{N} a_{il}x_{il,j} + b) \geqslant 1 - \delta_j, j = \overline{1,N} \\ \delta_j \geqslant 0, j = \overline{1,N} \end{cases} \tag{3}$$

Помимо возможности использования в качестве модальностей функций парного сравнения, RVM (3) отличается от классического критерия Бишопа и Типпинга в двух аспектах. Во-первых, критерий обучения выпуклый, следовательно, гарантированно имеет единственное решение. Во-вторых, он содержит структурный параметр μ, при помощи которого можно изменять селективность. Подходящее значение μ может быть выбрано, например, при помощи процедуры кросс-валидации.

Этот же критерий используется в данной работе и в задачах восстановления числовых зависимостей. Для нормализованной и центрированной выборки критерий, предложенный Зоу и Хасти, выглядит следующим образом:

$$\sum_{i=1}^{n}(\beta a_i^2 + \mu|a_i|) + \sum_{j=1}^{N}(y_j - \sum_{i=1}^{n} x_{ij}a_i)^2 \to \min_{\mathbf{a}} \tag{4}$$

Здесь $0 \leqslant \mu < +\infty, 0 \leqslant \beta < +\infty$.

В данной диссертации рассматривается обобщённый критерий для беспризнакового восстановления числовых зависимостей на основе нескольких конкурирующих функций парного сравнения согласно (2). Кроме того, нам будет более удобно исследовать процесс отбора вторичных признаков, записав критерий через остатки аппроксимации целевой переменной:

$$\begin{cases} \sum\limits_{i=1}^{n}(\beta a_i^2 + \mu|a_i|) + \sum\limits_{j=1}^{N}\delta_j^2 \to \min\limits_{\mathbf{a},\delta_1,\ldots,\delta_N} \\ \delta_j = y_j - \sum\limits_{i=1}^{n} x_{ij}a_i, j = \overline{1,N} \end{cases} \tag{5}$$

Во второй и третьей главах рассматриваются в деталях выпуклые критерии для задач распознавания образов (3) и восстановления числовых зависимостей (5), а также приводятся алгоритмы последовательной регуляризации и алгоритмы минимизации предложенных выпуклых критериев. Критерий (3), записанный для задачи распознавания образов, обнуляет часть коэффициентов

разделяющей гиперплоскости явным образом, тем самым производя отбор релевантных признаков. Это свойство данного критерия описывается следующей **теоремой**:

Теорема 1. *Оптимальная гиперплоскость* $(a_{il}, i = \overline{1,n}, l = \overline{1,N}, b)$ *определяется равенствами*

$$\begin{cases} \hat{a}_{il} = \frac{1}{2\beta}(s_{il} + \mu), s_{il} < -\mu, \\ \hat{a}_{il} = 0, -\mu \leqslant s_{il} \leqslant \mu, \\ \hat{a}_{il} = \frac{1}{2\beta}(s_{il} - \mu), s_{il} > \mu, \end{cases} \tag{6}$$

$$\hat{b} = -\frac{\sum_{j:0<\hat{\lambda}_j<1} \hat{\lambda}_j \sum_{l,i:\hat{\lambda}_l>0, \hat{a}_{il}\neq 0} \hat{a}_{il} x_{il,j} + \sum_{j:\hat{\lambda}_j=1} y_j}{\sum_{j:0<\hat{\lambda}_j<1} \hat{\lambda}_j}, \tag{7}$$

где

$$s_{il} = \sum_{j:\hat{\lambda}_j>0} y_j \hat{\lambda}_j x_{il,j},$$

$\{j : 0 < \hat{\lambda}_j < 1\}$ *и* $\{j : \hat{\lambda}_j = 1\}$ *– подмножества ненулевых решений* $\{j : \hat{\lambda}_j > 0\}$ *двойственной задачи выпуклого программирования:*

$$\begin{cases} W(\lambda_1, \ldots, \lambda_N | \mu) = \sum_{j=1}^N \lambda_j - \\ -\frac{1}{4\beta} \sum\limits_{i=1}^n \sum\limits_{l=1}^N \left\{ min \begin{bmatrix} \mu + \sum\limits_{j=1}^N y_j \lambda_j x_{il,j} \\ 0 \\ \mu - \sum\limits_{j=1}^N y_j \lambda_j x_{il,j} \end{bmatrix} \right\}^2 \to \max\limits_{\lambda_1,\ldots,\lambda_N}, \\ \sum_{j=1}^N y_j \lambda_j = 0, \\ 0 \leqslant \lambda_j \leqslant 1,\ j = \overline{1,N} \end{cases} \tag{8}$$

Аналогичная теорема для задачи восстановления числовых зависимостей:

Теорема 2. *Значения коэффициентов регрессии* $(a_{il}, i = \overline{1,n}, l = \overline{1,N}, b)$ *определяется равенствами*

$$\hat{a}_i(\delta_1, ..., \delta_N) = \begin{cases} \frac{1}{\beta}\left(\sum\limits_{j=1}^N \hat{\delta}_j x_{ij} + \mu\right), \sum\limits_{j=1}^N \hat{\delta}_j x_{ij} \leqslant -\mu/2, \\ 0, -\mu/2 < \sum\limits_{j=1}^N \hat{\delta}_j x_{ij} < \mu/2, \\ \frac{1}{\beta}\left(\sum\limits_{j=1}^N \hat{\delta}_j x_{ij} - \mu\right), \sum\limits_{j=1}^N \hat{\delta}_j x_{ij} \geqslant \mu/2, \end{cases} \tag{9}$$

где $\hat{\delta}_j, j = \overline{1,N}$ *– решение двойственной задачи квадратичного программирования:*

$$W(\delta) = \frac{1}{\beta}\sum_{i=1}^{n}\left\{\min\left[\frac{\mu}{2} + \mathbf{x}_i^T\delta, 0, \frac{\mu}{2} - \mathbf{x}_i^T\delta\right]\right\}^2 + (\delta - \mathbf{y})^T(\delta - \mathbf{y}) \to \min(\delta) \quad (10)$$

Доказательство обеих теорем базируется на анализе седловой точки функции Лагранжа, записанной для каждого из двух критериев. Главное свойство, даваемое приведёнными теоремами, заключается в том, что в решающем правиле часть коэффициентов оказывается нулевыми явно, а не из-за численных ошибок округления. Это даёт понимание, почему модуль приводит к тому, что часть коэффициентов обнуляется, а не оказывается бесконечно близкими к нулю.

Для численного решения задач (3) предлагается эффективный, хорошо масштабируемый на многопроцессорные системы итерационный алгоритм. Построение стандартного классификатора SVM сводится к решению задачи квадратичного программирования. Кроме того, на практике появляются неквадратичные задачи SVM, которые сводятся к решению выпуклой задачи оптимизации, но скорость работы стандартных методов становится узким горлышком при построении разделяющей гиперплоскости. Один из известных подходов к построению стандартного классификатора SVM – использовать эвристический алгоритм SMO. Во-первых, SMO заточен под решение именно квадратичной задачи, к которой сводится классический SVM, а значит неквадратичные задачи выпуклой оптимизации этим алгоритмом решать нельзя. Во-вторых, SMO делает большое количество итераций, поэтому на очень больших объёмах данных его ускорить не удастся.

Существующие способы, оптимизирующие скорость построения классификатора SVM, не годятся по следующим причинам.

- Классические методы решения SVM заточены под стандартный SVM, а именно его «квадратичность», поэтому не обобщаются на критерий произвольного вида.

- Способы класса SMO делают очень много «легковесных» итераций. Естественно, если выборка большая, такие алгоритмы будут работать очень долго. Задействовать мощности современных кластеров в случае с SMO либо с другим аналогичным алгоритмом, делающим много итераций, невыгодно.

Решение критерия с квадратично-модульной регуляризацией аналитически выписать не удаётся, во-первых, из-за модуля, а, во-вторых, из-за линейных ограничений. Поэтому для поиска точки оптимума используется двойственная задача и итерационные методы решения задач математического программирования. Двойственная задача к критерию с квадратично-модульной регуляризацией является стандартной задачей квадратичного программирования с линейными ограничениями. Для её решения можно применять следующие классические методы численной условной оптимизации:

- Метод проекции градиента

- Метод условного градиента
- Метод внутренней точки (метод штрафных функций)

Каждый из перечисленных методов обладает своими преимуществами и недостатками. Нас в первую очередь интересует скорость работы этих методов и требуемая для их работы память. Метод проекции градиента и метод условного градиента требуют линейных расходов по памяти в зависимости от размерности задачи. Однако, их слабая сторона - количество итераций, которые эти методы делают. Очень часто даже в задачах небольшой размерности ($N \sim 100$) эти методы выполняют тысячи итераций. Естественно, что как бы отдельная итерация ни была ускорена, все равно принципиально большой скорости достигнуть не получится в силу того, что этих итераций выполняется очень много.

Метод внутренней точки в этом смысле обладает гораздо более высокими показателями скорости. Даже на больших объёмах данных ($N \sim 1000$) этот метод выполняет десятки итераций. Однако, каждая итерация вычислительно оказывается тяжёлой. Связано это с тем, что метод внутренней точки сводит исходную задачу математического программирования к задаче безусловной оптимизации, добавляя к целевой функции штрафные функции, штрафующие выход переменных оптимизации за пределы области допустимых ограничений. Сама задача безусловной оптимизации решается методом Ньютона, что означает, что на каждой итерации нужно перемножить матрицы размеров $N \times N$ и решить систему линейных уравнений размеров $N \times N$. Первая из этих операций очень хорошо параллелится. Вторая операция тоже может быть распараллелена, но менее эффективно. Метод внутренней точки требует много памяти для хранения матриц размеров $N \times N$, то есть потребление памяти в этом случае растет квадратично от размерности задачи.

В данной работе предлагается подход к реализации обучения дважды регуляризованной машины RVM, основанный на методе внутренней точки. Во-первых, этим методом можно решать произвольную задачу выпуклой оптимизации, целевая функция которой дважды непрерывно дифференцируема. Во-вторых, метод в среднем делает существенно меньшее количество итераций по сравнению с SMO и имеет отличный потенциал для ускорения через распараллеливание. Ясно, что квадратичное использование памяти делает этот метод неприменимым при работе с очень большим количеством данных. В рамках текущей работы авторы не столкнулись с задачами такой размерности, при которой становится невозможным применять метод внутренней точки.

Для численного поиска решения критерия (5) в задачах восстановления числовых зависимостей предлагаются эффективные алгоритмы регуляризации.

Идея алгоритма регуляризации впервые была сформулирована для критерия Лассо($\beta = 0$) и затем расширена на случай Elastic Net($\beta > 0$). Мотивацией послужило предположение, что поиск зависимости вектора коэффициентов

регрессии от параметра селективности $\hat{\mathbf{a}}_{\beta\mu} = \hat{\mathbf{a}}_\beta(\mu)$ в один проход сразу для всех значений селективности должен быть вычислительно гораздо эффективнее, чем независимое вычисление искомого вектора коэффициентов регрессии для нескольких отдельных значений селективности.

Пусть коэффициент $\beta > 0$ взят достаточно малым, чтобы гарантировать строгую выпуклость целевой функции Elastic Net (1). В основе теоретического обоснования алгоритма регуляризации для всех значений параметра селективности $\mu > 0$ лежит математический факт, что зависимость $\hat{\mathbf{a}}_\beta(\mu)$, определённая обучающим критерием Elastic Net - непрерывная, кусочно-линейная функция, имеющая конечное количество точек бифуркации. Будем рассматривать эти точки как уменьшающуюся последовательность $\big(\mu_{max} = \mu_0 > \mu_1 > , ..., > \mu_M = 0\big)$. Можно доказать, что количество нетривиальных точек M удовлетворяет неравенству $n \leqslant M \leqslant 2^n$, где $n = |\mathbb{I}|$ - количество признаков.

Это значит, что в терминах теоретического подхода для каждой обучающей совокупности убывающая последовательность точек бифуркации $\mu_{max} \to \mu_k \to 0$ задаёт соответствующую дискретную последовательность разбиений множества признаков P_k, начиная с разбиения, в котором все признаки отсутствуют $\hat{\mathbb{I}}_0 = \hat{\mathbb{I}}_0^- \bigcup \hat{\mathbb{I}}_0^+ = \varnothing$, и заканчивая разбиением, включающим в себя все признаки $\hat{\mathbb{I}}_M = \mathbb{I}$. Между двумя соседними точками бифуркации разбиение остаётся неизменным $\mu_{k-1} > \mu > \mu_k$, но следует отметить, что, вообще говоря, размер множества активных признаков $|\hat{\mathbb{I}}_k|$ не является монотонно возрастающей функцией от значения параметра селективности μ.

Теоретически всего точек бифуркации может быть 2^n. Однако такое число чрезмерно велико для применения предлагаемого алгоритма Машины Релевантных Объектов, базирующегося на принципе отбора признаков. В нашем случае количество вторичных признаков $n = |\mathbb{I}| = mN$ превышает размер обучающей совокупности $N = |\mathbb{J}|$ в такое же количество раз, сколько задано функций парного сравнения объектов $m = |\mathbb{K}|$. Поэтому мы будем использовать «урезанный» алгоритм регуляризации.

Окончательным результатом алгоритма регуляризации является последовательность подмножеств активных признаков $\hat{\mathbb{I}}_k$, $k = 0, 1, ..., M$. Мощность этих подмножеств $|\hat{\mathbb{I}}_k|$ в общем случае возрастает от 0 до полного количества признаков $n = mN$, но её зависимость от параметра селективности необязательно монотонная.

Четвертая глава посвящена численной реализации алгоритмов, изложенных во второй и третьей главах. Алгоритмы регуляризации и алгоритмы решения выпуклой задачи (3) устроены так, что это итерационные алгоритмы, основное время работы которых заключается в выполнении на каждой итерации операций умножения и решения системы линейных уравнений для матриц размеров $N \times N$, где N - количество объектов обучающей совокупности.

Кроме стандартного алгоритма умножения матриц $N \times N$ за время $O(N^3)$, известны также алгоритмы, работающие быстрее:

1. Алгоритм Штрассена умножает две матрицы размеров $N \times N$ за $O(N^{2.807})$

2. Алгоритм Копернika-Винограда , имеющий сложность умножения $O(N^{2.3755})$

3. Алгоритм Вильямс, имеющий сложность $O(N^{2.3727})$

Представим, что у нас в распоряжении очень простая вычислительная система - видеокарта NVidia Geforce 310M, имеющая 16 GPU. Оценим, каких размеров должна быть матрица, чтобы алгоритм Штрассена дал выигрыш по сравнению с распараллеливанием на 16 ядер. Ответ прост: $N = 16^{\frac{1}{3-2.807}} \approx 1000000$. В оперативной памяти такая матрица занимала бы $\frac{8 \times 1000000^2}{1024^2} \approx 8$ терабайт, что на текущий момент технически невозможно принципиально. Ясно, что при переходе к более мощным вычислительным системам выигрыш в скорости будет только увеличиваться при неизменных размерах матрицы. Что касается алгоритма Вильямс и алгоритма Копернika-Винограда, то эти алгоритмы имеют очень большую константу сложности, поэтому выигрывают у современных алгоритмов на матрицах, размеры которых превосходят современные технические возможности компьютеров. Поэтому в рамках данной работы параллельное умножение матриц на видеокарте - лучшее решение из всех возможных, исходя из количества арифметических операций. Если рассмотреть детали технической реализации, то необходимо учесть тот факт, что матрицу, загруженную в оперативную память компьютера, необходимо 1 раз скопировать в оперативную память видеокарты, что по порядку величины занимает $O(N^2)$ операций. Учитывая современные скорости передачи информации CPU $\rightarrow$ GPU и GPU $\rightarrow$ CPU, даже для матрицы 10000×10000 подобная операция копирования займет меньше секунды, что в разы меньше времени, которое тратится на операции умножения матриц на GPU. В рамках данной работы выполнено две реализации алгоритмов умножения матриц: для однопроцессорной машины(CPU) и для видеокарты(GPU). В качестве фреймворка, на базе которого выполнялась реализация алгоритмов, был взят OpenCL. Это удобный современный фреймворк, сочетающий в себе возможность гибкой разработки на C++ и реализованный практически подо все современные видеокарты, в частности карты NVidia. Видно, что выигрыш при перемножении матриц больших размеров доходит до 25 раз. Это связано не только с тем, что на видеокарте большее количество ядер, чем на центральном процессоре, но ещё и с тем, что планировщик параллельных вычислений на видеокарте работает лучше и оптимальнее распределяет нагрузку, чем аналогичный планировщик на CPU.

Существующие способы решения системы линейных уравнений, матрицы которых не являются разреженными, делятся на точные(прямые) методы и итерационные методы. В работе используется точный метод решения системы линейных уравнений на базе разложения Холецкого.

В пятой главе приведены результаты численных экспериментов применения предложенных алгоритмов для модельных экспериментов и в реальных задачах прогнозирования вторичной структуры белков и определения намагниченности в сверхпроводниках.

Рис. 1: Среднее время работы двух разных реализаций алгоритма умножения матриц – на видеокарте и на однопроцессорной машине – в зависимости от размера умножаемых матриц.

Для предсказания вторичной структуры белка на основе его фрагментов мы применили Машину Релевантных Объектов, описанную в данной работе. Мы ограничились рассмотрением задачи определения strand'ов во вторичной структуре белков, которая, как показывает практика, представляет собой наиболее проблематичную часть вторичной структуры. Целью данной работы является скорее исследование эффективности RVM к проблеме оконного предсказания вторичной структуры, чем установление нового рекорда точности. Тем не менее, эксперименты на RS126 показали точность около 75% в определении strand'ов.

Для проверки способности Машины Релевантных Объектов выбирать наиболее подходящие функции сравнения, мы исследовали вместе два принципа сравнения – принцип, основанный на позициях аминокислот во фрагменте, и новый подход, основанный на разложении Фурье первичной и вторичной структуры белка.

Пусть $A = \{\alpha^1, \cdots, \alpha^{20}\}$ алфавит на множестве аминокислот. Для каждой позиции $-T \leqslant \tau \leqslant T$ в окне $\omega = (\alpha_\tau, -T \leqslant \tau \leqslant T)$ и каждой из 20 аминокислот определен бинарный признак $z_{\tau k}(\omega) = 1$, если $\alpha_\tau = \alpha^k$, и $z_{\tau k}(\omega) = 0$, если $\alpha_\tau \neq \alpha^k$.

$$
\begin{aligned}
S_1(\omega', \omega'') &= \sum_{\tau=-T}^{T} \sum_{k=1}^{20} z_{\tau k}(\omega') z_{\tau k}(\omega''), \\
S_2(\omega', \omega'') &= \exp\left\{-\gamma \left[z_{\tau k}(\omega') - z_{\tau k}(\omega'')\right]^2\right\}, \\
S_3(\omega', \omega'') &= \sum_{\tau=-T}^{T} \sum_{k=1}^{20} \left|z_{\tau k}(\omega') - z_{\tau k}(\omega'')\right|.
\end{aligned}
\tag{11}
$$

Все признаки образуют бинарный вектор $z(w) = z_{\tau k}(\omega), -T \leqslant \tau \leqslant T, k = 1, \ldots, 20)$. Мы исследовали три вышеозначенные функции сравнения, основанные на таком признаковом описании объектов. В статьях, посвящённых сравнению белков на базе фрагментов, показано, что позиционный принцип сравнения работает лучше, когда применяется к сравнительно небольшим длинам окон, поэтому использовалась рекомендованная длина окна $2T + 1 = 13$.

Суть сравнения двух аминокислотных фрагментов (ω', ω'') на базе разложения Фурье заключается в использовании вектора признаков в рамках одного метода сравнения $S(\omega', \omega'') = S(\mathbf{f}(\omega'), \mathbf{f}(\omega''))$. Мы исследовали три функции:

$$\begin{aligned} S_4(\omega', \omega'') &= \mathbf{f}^T(\omega')\mathbf{f}(\omega''), \\ S_5(\omega', \omega'') &= \exp\left\{-\gamma||\mathbf{f}(\omega') - \mathbf{f}(\omega'')||^2\right\}, \\ S_6(\omega', \omega'') &= \sum_{k=1}^{20}\sum_{l=0}^{4} |f_{kl}(\omega') - f_{kl}(\omega'')|. \end{aligned} \tag{12}$$

В качестве обучающей выборки использовался стандартный набор белков RS126. Этот набор содержит 126 белков, среди которых схожестью менее 25% обладают все белки длиной более 80 аминокислот. Точность классификации оценивалась для разных уровней селективности аминокислотных фрагментов и разных функций сравнения этих фрагментов. В общей сложности из белков набора RS126 получилось $|\Omega| = 19075$ фрагментов $\omega \in \Omega$ длиной $2T + 1 = 35$. Каждый фрагмент отнесен к классу $y = \pm 1$, согласно тому, является ли его центральный элемент стрендом или не является. Мы выполнили 4 эксперимента на данном наборе данных.

В каждом эксперименте множество всех окон Ω было случайным образом разбито на обучение $\Omega_{tr} \in \Omega$ размера $N = |\Omega_{tr}| = 1600$ и контроль $\Omega_{test} = \Omega \setminus \Omega_{tr}$ размера $|\Omega_{test}| = 17475$.

Множество функций сравнения формируется из (11), (12), т.е. их $n = 6$. Функции сравнения, основанные на методе Фурье (12), учитывают $2T+1 = 35$ аминокислот каждого фрагмента, а сравнение по позициям (11), применяемое к более коротким фрагментам длины $2T + 1 = 13$, игнорирует 11 аминокислот с обеих сторон фрагментов длины 35.

Каждый из четырёх экспериментов был устроен следующим образом. Мультимодальный RVM обучался на совокупности $\Omega_{tr}, N = 1600$ для всех $\mu \in \{0, 0.3, 0.5, 0.6, 0.8, 0.9999, 0.99999(\mu \to 1)\}$. Максимальное значение параметра $\mu \to +\infty$ в данном эксперименте намеренно заменено на $\mu \to 1$, так как экспериментально было определено, что при $\mu = 1$ все признаки исключаются из решающего правила. Таким образом, обучение мультимодальной RVM запускалось $4 \times 7 = 28$ раз. Это серьезный объём вычислений для настольного компьютера.

Результат каждого обучения мультимодального RVM с конкретным значением μ есть подмножество вторичных релевантных признаков $\hat{F} = \{i, j :$

$\hat{a}_{ij}(\mu) \neq 0\} \subseteq F$ и значений параметров $\left(a_{ij}(\mu) \in \hat{F}, b(\mu)\right)$ разделяющей гиперплоскости. Наиболее важное значение имеют множества релевантных объектов(аминокислотных фрагментов обучающей совокупности)$\hat{J}(\mu) = \{j : \exists i(a_{ij} \neq 0)\} \subseteq \{\overline{1,N}\}$ и релевантных модальностей $\hat{I}(\mu) = \{i : \exists j(a_{ij} \neq 0)\} \subseteq I = \{\overline{1,n}\}$. Обозначим мощности этих двух множеств $\hat{N}(\mu) = |\hat{J}(\mu)| \leqslant N$ и $\hat{n}(\mu) = |\hat{I}(\mu)| \leqslant n$. Параметр β был определен на основании процедуры кроссвалидации и составил $\beta = 0.001$.

Оставшееся множество $\Omega_{test} = \Omega \setminus \Omega_{tr}$ аминокислотных фрагментов было случайно разбито на 10 подмножеств для контроля, после чего мы посчитали точность распознавания вторичной структуры {стренд} против {нестренд} для каждого из них. Окончательная точность для каждого из значений селективности характеризуется двумя числами: мат. ожиданием $Acc(\mu)$ и дисперсией $\sigma(\mu)$. Доверительный интервал для точности классификации мы оценили как $Acc(\mu) \pm 2\sigma(\mu)$.

Рисунок (2) демонстрирует зависимость точности предсказания $Acc(\mu)$ и количества релевантных аминокислотных фрагментов, участвующих в решающем правиле $\hat{N}(\mu)$, от уровня селективности μ. Из рисунка видно, что во всех экспериментах наилучшая точность классификации 75.5% достигается при $\mu = 0$, т.е. когда все 1600 аминокислотных фрагментов, образующих обучающую совокупность, и все 6 функций сравнения участвуют в разделяющей гиперплоскости. Последняя соответствует пространству вторичных признаков размерностью $nN = 9600$, получаемых из фрагмента $\omega = (\alpha_\tau \in \mathbb{A}, -T \leqslant \tau \leqslant T)$. Особенно интересно, что не наблюдается эффекта переобучения после построения разделяющей гиперплоскости в линейном пространстве векторов вторичных признаков $\mathbf{x}(\omega) = \left(x_{ij}(\omega), i = \overline{1,6}, j = \overline{1,1600}\right) \in \mathbb{R}^{9600}$, чья размерность в 6 раз превышает количество объектов обучающей совокупности.

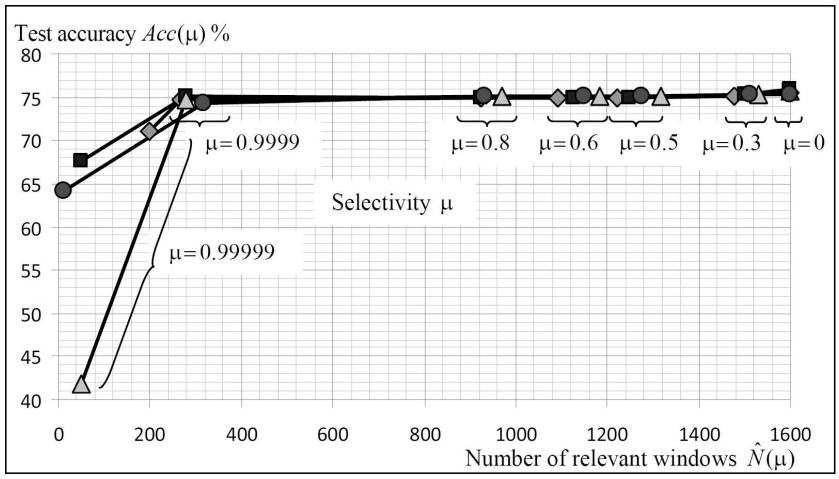

Рис. 2: Экспериментальная зависимость количества релевантных аминокислотных фрагментов $\hat{N}$ и точности распознавания стрендов на контроле Acc от уровня селективности μ.

С ростом μ уменьшается количество релевантных фрагментов $\hat{N}(\mu)$ и

количество релевантных функций сравнения $\hat{n}(\mu)$, формирующих вторичные признаки аминокислотных фрагментов. Точность остаётся практически на одном уровне во всех независимых экспериментах вплоть до уровня селективности $\mu = 0.9999$, когда осталось порядка 300 релевантных аминокислотных фрагментов из 1600 и только $\hat{n} = 3$ функции сравнения. Уменьшение точности по отношению к её уровню для $\mu = 0$ не превосходит 1%.

При дальнейшем увеличении $\mu > 0.9999$ происходит значительное снижение точности распознавания на всех обучающих совокупностях.

Кроме задачи распознавания образов, в данной работе эффективность Машины Релевантных Объектов проверена на задаче восстановления числовой зависимости, которая заключается в моделировании намагниченности в сверхпроводниках. Данные для этой задачи предоставлены Национальным Технологическим Институтом Стандартов(National Institute of Standards and Technology) `http://www.nist.gov/index.html`.

Набор данных состоит из объектов $\omega \in \Omega$, представленных одним наблюдаемым признаком $\mathbf{z}(\omega) = z_1(\omega) \in \mathbb{R}$ и одним скрытым признаком $y(\omega) \in \mathbb{R}$.

Скрытый признак – это измеренная намагниченность сверхпроводника, наблюдаемый признак – это логарифм от времени, прошедшего с момента старта эксперимента до момента его завершения. Истинная зависимость между скрытым и наблюдаемым признаком описывается $y = -2000(50 + z_1)^{-10/9}$.

Набор данных содержит $|\Omega| = 154$ объекта, которые случайно разбиты на $N = 75$ объектов для обучения и $N_{test} = 79$ объекта для контроля. На Рис. (3) показана заданная скрытая функция $y^*(\mathbf{z})$, которую нужно оценить, и обучающая совокупность $\{(\mathbf{z}_j, y_j), j = 1, ..., N\}$.

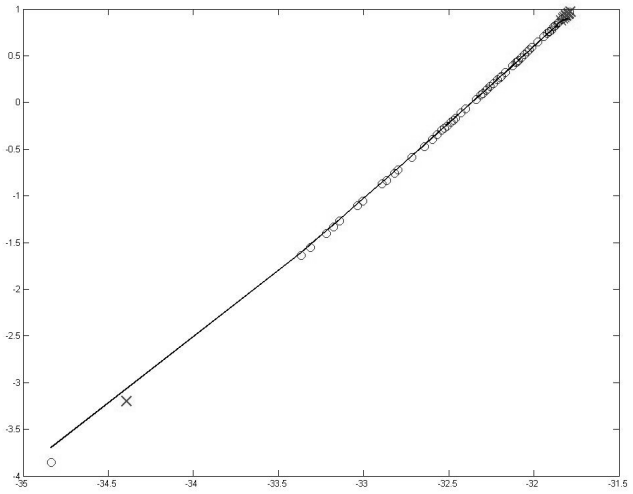

Рис. 3: Обучающая совокупность $N = 75$ и восстановленная по ней зависимость $y^*(\mathbf{z})$. Маркеры $\odot$ и $\times$ обозначают объекты обучающей совокупности $\mathbf{z}_j$ и, соответственно, релевантные объекты.

Далее мы применили к обучающей совокупности предложенную в работе Машину Релевантных Объектов с отбором модальностей для фиксированного значения параметра $\beta = 15$ и для $m = 4$ функций парного сравнения:

$$
\begin{array}{l}
S_1(\omega',\omega'') = S_1(\mathbf{z}',\mathbf{z}'') = (1+|z_1'-z_1''|)^{-10/9}, \\
S_2(\omega',\omega'') = S_2(\mathbf{z}',\mathbf{z}'') = \exp(-1.5(z_1'-z'')^2), \\
S_3(\omega',\omega'') = \exp(-1.5*|z_1'-z_1''|)/(1+(z_1'+z_1'')^2), \\
S_4(\omega',\omega'') = S_4(\mathbf{z}',\mathbf{z}'') = \exp(-1.5|z_1'-z_1''|)
\end{array} \tag{13}
$$

Особенность скрытой заданной функции указывает на то, что алгоритм должен отобрать только первую функцию $S_1(\mathbf{z}',\mathbf{z}'')$ как единственную адекватную функцию искомой зависимости, а в качестве релевантных объектов отобрать небольшое количество объектов обучающей совокупности.

Таким образом, мы имеем полный набор $\mathbb{I} = \{i = (kj)\}$ из $n = mN = 300$ вторичных признаков, сформированный из $m = 4$ модальностей и $N = 75$ объектов обучения. Для подобной экспериментальной структуры максимальное значение параметра селективности равняется $\mu_{max} \approx 94.5$.

Применение алгоритма регуляризации привело к тому, что лучшее значение селективности равно $\mu = 186.4$, давая минимальное среднее значение ошибки leave-one-out $Q^{LOO}_{\hat{\mathbb{I}}} = 0.000695$. Средний квадрат ошибки на контроле оказался немного выше: $Q^{test}_{\hat{\mathbb{I}}} = 0.000805$. Обе оценки хорошо согласуются со значением квадрата отклонения истинной зависимости от наблюдаемых значений скрытой переменной $Var(y) = 0.00052$.

Соответствующее подмножество релевантных вторичных признаков $\hat{\mathbb{I}} = \{(kj)\}$ содержит 11 элементов. Все эти элементы были сформированы разными релевантными объектами $\hat{\mathbb{J}} = \{j\} \subset \mathbb{J}$, связанными с одной и той же функцией парного сравнения $S_1(\mathbf{z}',\mathbf{z}'')$, как и ожидалось.

ЗАКЛЮЧЕНИЕ

Основные результаты работы:

1. Разработан комплекс выпуклых критериев отбора релевантных подмножеств объектов обучающей совокупности и функций парного сравнения с регулируемой селективностью в задачах обучения распознаванию образов и восстановления числовых зависимостей.

2. Разработан комплекс итерационных алгоритмов решения задач обучения распознаванию образов и восстановления числовых зависимостей с заданной селективностью отбора релевантных объектов и релевантных функций парного сравнения.

3. Разработан комплекс многопроцессорных алгоритмов реализации отдельной итерации обучения распознаванию образов и восстановления числовых зависимостей с заданной селективностью

4. Разработаны последовательные алгоритмы выбора уровня селективности в задачах обучения на основе парного сравнения объектов.

5. Создан программно-алгоритмический комплекс, реализующий алгоритмы селективного комбинирования разнородных представлений объектов, хорошо масштабирующихся на большое число процессоров.

6. Экспериментально исследованы разработанные алгоритмы обучения распознаванию образов и восстановления числовых зависимостей на основе селективного комбинирования функций парного сравнения объектов.

Перспективы дальнейшей разработки. В работе предлагается такая концепция отбора релевантных объектов и релевантных функций парного сравнения, в соответствии с которой степень отбора задаётся общей и для отбора релевантных функций парного сравнения и для отбора релевантных объектов. На практике возникают задачи, когда уровень селективности нужно сделать разным для отбора релевантных объектов и релевантных функций парного сравнения. Предложенный в работе класс критериев этого сделать не позволяет – требуется дополнительное исследование, чтобы решить эту задачу. В работе разработан последовательный алгоритмы отбора селективности для задач восстановления числовых зависимостей, однако для задач распознавания образов – это открытый вопрос, который нуждается в исследованиях. Наконец, в работе предложен комплекс итерационных алгоритмов решения задач обучения распознаванию образов и восстановления числовых зависимостей с заданной селективностью отбора релевантных объектов и релевантных функций парного сравнения, однако эти алгоритмы имеют квадратичный рост потребляемой ими памяти в зависимости от объёма данных, что ограничивает применение предлагаемых в работе алгоритмов на сверхбольших объёмах данных. Если возникнет потребность в таких задачах, потребуется дополнительное исследование, включающее в себя улучшение предложенных в данной работе численных методов.

Основные положения диссертации опубликованы в следующих работах

[1] Выпуклые критерии релевантных векторов для сокращения размерности представления объектов в беспризнаковом распознавании образов / Олег Середин, Вадим Моттль, Александр Татарчук, Николай Разин // Сборник трудов ИОИ-9. — 2012. — С. 50 – 53.

[2] Применение мультимодальной rvm в задаче распознавания вторичной структуры белков / Николай Разин, Дмитрий Сунгуров, Вадим Моттль и др. // Сборник трудов ИОИ-9. — 2012. — С. 585 —- 588.

[3] Convex support and relevance vector machines for selective multimodal pattern recognition / Oleg Seredin, Vadim Mottl, Alexander Tatarchuk et al. // ICPR-2012 proceedings. — 2012.

[4] **Multi-Modal Relevance Vector Machines for protein secondary structure prediction** / Nikolay Razin, Dmitry Sungurov, Vadim Mottl et al. // PRIB-2012 proceedings. — 2012. — P. 153 —- 165.

[5] **Выпуклые критерии релевантных векторов для сокращения размерности описания объектов, представленных парными отношениями** / О. С. Середин, В. В. Моттль, А. И. Татарчук, Н. А. Разин // Известия Тульский Государственный Университет. — 2013. — Т. 1. — С. 165 – 176.

[6] Применение Машины Релевантных Объектов в задачах восстановления числовых зависимостей / Разин Н. А., Черноусова Е. О., Красоткина О. В., Моттль В. В // Машинное обучение и анализ данных. — 2013. — Т. 1, № 5. — С. 641 – 652.

[7] Разин Н. А., Моттль В. В. Численная реализация алгоритмов селективного комбинирования разнородных представлений объектов в задачах распознавания образов // Машинное обучение и анализ данных. — 2013. — Т. 1, № 6. — С. 734 – 750.

Жирным шрифтом выделены публикации в журналах, рекомендованных ВАК. Все приведённые в статьях результаты исследований получены лично автором. Все статьи, кроме [3], [4], написаны лично автором с последующим участием соавторов в редактировании текста.

Разин Николай Алексеевич

ВЫПУКЛЫЕ КРИТЕРИИ И ПАРАЛЛЕЛИЗУЕМЫЕ АЛГОРИТМЫ СЕЛЕКТИВНОГО КОМБИНИРОВАНИЯ РАЗНОРОДНЫХ ПРЕДСТАВЛЕНИЙ ОБЪЕКТОВ В ЗАДАЧАХ ВОССТАНОВЛЕНИЯ ЗАВИСИМОСТЕЙ ПО ЭМПИРИЧЕСКИМ ДАННЫМ

АВТОРЕФЕРАТ

Подписано в печать 11.11.2013.
Формат $60\times84\ {}^{1}/_{16}$. Усл. печ. л. 1,0. Тираж 100 экз. Заказ № 338.
Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Московский физико-технический институт (государственный университет)»
Отдел оперативной полиграфии «Физтех-полиграф»
141700, Московская обл., г. Долгопрудный, Институтский пер., д. 9.