Гуз Иван Сергеевич

Комбинаторные оценки полного скользящего контроля и методы обучения монотонных классификаторов

05.13.17

физико-математические науки

Д 002.017.02

Учреждение Российской академии наук Вычислительный центр им. А.А. Дородницына РАН

119333, г. Москва, ул. Вавилова, д.40.

Тел.: (499) 135-01-89

E-mail: aspiran@ccas.ru

Предполагаемая дата защиты – 8 декабря 2011 года

На правах рукописи

Гуз Иван Сергеевич

Комбинаторные оценки полного скользящего контроля и методы обучения монотонных классификаторов

Специальность 05.13.17 – теоретические основы информатики

Автореферат

диссертации на соискание учёной степени
кандидата физико-математических наук

Москва – 2011

Работа выполнена на кафедре интеллектуальных систем
Московского физико-технического института
(государственного университета).

Научный руководитель:

доктор физико-математических наук,
ВОРОНЦОВ Константин Вячеславович.

Официальные оппоненты:

академик РАН,
доктор физико-математических наук, профессор
МАТРОСОВ Виктор Леонидович,

кандидат физико-математических наук
РОМАНОВ Михаил Юрьевич.

Ведущая организация:

Московский государственный университет имени М.В. Ломоносова, факультет ВМК.

Защита диссертации состоится «____»_____ 2011 г. в ____ часов на заседании диссертационного совета Д 002.017.02 в Учреждении Российской академии наук Вычислительного центра им. А.А. Дородницына РАН по адресу: 119333, г. Москва, ул. Вавилова, д.40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан «____»_____2011 г.

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

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

Проблема синтеза алгоритмов на основе обучения по конечным выборкам прецедентов и изучения их качества на всем множестве прецедентов является одним из важнейших вопросов теории обучения по прецедентам. В качестве прецедентов рассматриваются пары: объект, описанный набором признаков, и класс, к которому принадлежит объект. Задача классификации состоит в том, чтобы на основании известного конечного множества прецедентов научиться определять априори неизвестную принадлежность объектов к классам. Обучение или настройка параметров алгоритма на обучающей выборке происходит путем решения задачи численной оптимизации. Практика показала, что при решении прикладных задач классификации очень часто возникает ситуация, когда ни один из существующих алгоритмов в отдельности не решает задачу с достаточным качеством. В таких случаях пытаются учесть сильные стороны каждого отдельного алгоритма за счет построения из них некоторой композиции.

В работе рассматривается проблема повышения качества классификации при помощи построения алгоритмических композиций для задач, описываемых некоторой выборкой объектов, где каждый объект принадлежит одному из двух непересекающихся классов, –1 и +1. Предполагается, что существует набор базовых алгоритмов, причем каждый базовый алгоритм определяет для каждого объекта не только его класс, но и оценку принадлежности классу +1. Этим свойством обладают многие известные алгоритмы классификации, например, байесовские классификаторы, нейронные сети, логистическая регрессия, дерево решений CART и другие. В байесовских классификаторах оценка принадлежности интерпретируется как апостериорная вероятность того, что объект принадлежит классу +1. Однако в данной работе никаких предположений о вероятностной природе данных не делается, и оценки принадлежности интерпретируются в более широ-

ком смысле. Чем больше оценка, тем с большей уверенностью можно утверждать, что объект принадлежит классу +1.

В качестве алгоритмической композиции в работе рассматривается монотонная корректирующая операция, которая является монотонным классификатором в пространстве оценок принадлежности. Использование монотонного классификатора оправдано естественным требованием, что если для одного объекта оценки принадлежности не меньше, чем для другого, то и оценка принадлежности первого объекта, рассчитанная с помощью композиции, должна быть не меньше, чем для второго. Монотонные корректирующие операции образуют более широкое семейство по сравнению с выпуклыми (линейными с неотрицательными коэффициентами), используемыми в методах голосования, в частности, в бустинге. Это позволяет точнее настраиваться на данные и использовать существенно меньшее число базовых алгоритмов, но одновременно с этим, повышает риск переобучения.

Уменьшение риска переобучения гарантирует доказанная в работе верхняя оценка полного скользящего контроля для семейства монотонных классификаторов, называемая **гибридной**, учитывающая как свойства задачи, так и особенности используемого семейства. Использование этой оценки в качестве функционала качества монотонной композиции позволяет уменьшить риск переобучения и повысить совокупное качество классификации.

Цели диссертационной работы

Целями диссертации являются:

1) Вывод комбинаторных оценок полного скользящего контроля для семейства монотонных алгоритмов
2) Создание на основе этих оценок методов построения алгоритмических композиций с монотонной корректирующей операцией, уменьшающих риск переобучения.

Научная новизна работы

1. Для одномерных обучающих выборок предложена вычислительно эффективная процедура расчета верхней и нижней оценок полного скользящего контроля при обучении монотонных алгоритмов на объектах с весами. Предложен метод фильтрации шумовых объектов на основе расчета полученной верхней оценки полного скользящего контроля.

2. Для многомерных обучающих выборок получена новая комбинаторная оценка полного скользящего контроля для семейства монотонных алгоритмов, учитывающая как свойства задачи, так и особенности используемого семейства. Основное достоинство полученной оценки в том, что она применима к любым обучающим выборкам, причем для монотонных обучающих выборок значение совпадает с фактическим значением функционала полного скользящего контроля. Предложенная оценка на обучающих выборках размерности больше двух точнее оценок, полученных ранее другими авторами.

3. Предложен новый метод построения монотонной композиций алгоритмов, основанный на уменьшении оценки полного скользящего контроля для монотонной корректирующей операции, существенно уменьшающий риск переобучения и повышающий качество всей композиции.

Методы исследования

Оценки полного скользящего контроля для класса монотонных алгоритмов получены на основании: комбинаторики, теории графов. Точность оценок подтверждена вычислительными экспериментами на модельных задачах классификации.

Для построения алгоритмической композиции с монотонной корректирующей операцией предполагается, что каждый из базовых алгоритмов способен обучаться на объектах обучающей выборки с весами. Чем больше вес объекта, тем точнее должен настраиваться на него базовый алгоритм. На текущий момент известно большое количество алгоритмов, допускающих такой способ обучения. Чтобы уменьшить риск переобучения в диссертации используются методы дискретной оптимизации для выбора весов объектов, с которыми следует обучать

каждый базовый алгоритм. Обучение проводится таким образом, чтобы оценка полного скользящего контроля для всей монотонной композиции была минимальна. Проведенные вычислительные эксперименты показали, что предложенный метод построения монотонной композиции действительно повышает совокупное качество классификации за счет уменьшения переобучения.

Положения, выносимые на защиту:

1. Комбинаторная оценка полного скользящего контроля для семейства монотонных алгоритмов, учитывающая как свойства задачи, так и свойства самого семейства монотонных алгоритмов, и результаты экспериментов, доказывающие точность этой оценки.
2. Метод построения оптимального монотонного классификатора, минимизирующего эмпирический риск.
3. Метод построения монотонной композиции базовых алгоритмов, минимизирующий полученную оценку полного скользящего контроля для уменьшения переобучения и повышения качества всей композиции.
4. Результаты вычислительных экспериментов, свидетельствующие о том, что разработанный метод построения монотонной композиции повышает качество класификации по сравнению с отдельными базовыми алгоритмами, а также некоторыми другими известными методами построения композиций.

Теоретическая ценность работы

В диссертации развивается теория надёжности обучения по прецедентам К. В. Воронцова: доказывается возможность построения качественных алгоритмических композиций без каких-либо априорных предположений о вероятностной природе распределения объектов обучения.

Практическая ценность работы

Метод построения монотонных алгоритмических композиций, предложенный в работе, может быть применен для повышения качества классификации в

тех случаях, когда каждый из базовых алгоритмов по отдельности не решает задачу с достаточным качеством. Причем в качестве базовых могут быть использованы любые алгоритмы, допускающие метод обучения на объектах с весами. Отличительной особенностью предложенного метода является отсутствие эффекта переобучения при увеличении количества базовых алгоритмов.

Данный метод также может быть применен при построении монотонного классификатора в задачах бинарной классификации в медицине, экономике, биологии и других областях. В большинстве таких задач предполагается монотонная зависимость между числовым признаковым описанием объектов и их принадлежностью к целевому классу. Применение предложенного метода позволяет учесть априорные монотонные ограничения и избежать переобучения.

Апробация и публикации

По теме диссертации опубликовано 7 работ, в том числе две работы [2,3] — в изданиях из списка, рекомендованного ВАК РФ. Результаты диссертационного исследования докладывались, обсуждались и получили одобрение специалистов на научных конференциях и семинарах:

- 15-я Всероссийская конференция «Математические методы распознавания образов», г. Петрозаводск, 2011 г. [1];
- 14-я Всероссийская конференция «Математические методы распознавания образов», г. Суздаль, 2009 г. [4];
- 7-я международная конференция «Интеллектуализация обработки информации», г. Симферополь, 2007 г. [7];
- 49-я научная конференция МФТИ, г. Москва, 2006 г. [8];
- Научные семинары отдела Интеллектуальных систем Вычислительного центра РАН и кафедры «Интеллектуальные системы» МФТИ, г. Москва, 2007-2011 г.г.

Структура диссертации

Диссертация состоит из введения, трех глав, заключения и списка использованных источников, включающего 43 наименования. Общий объем работы составляет 114 страниц.

КРАТКОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обоснована актуальность диссертационной работы, сформулированы цель и задачи исследования, описаны научная новизна полученных результатов, структура диссертации и кратко изложено содержание работы.

Глава 1. Монотонные классификаторы

1.1. Задача обучения по прецедентам. Пусть задано конечное множество $\mathbb{X} = \{x_1, x_1, \ldots, x_L\}$, состоящее из L *объектов*, в котором каждый объект x_i описывается вектором из n вещественных *признаков* $\{x_i^1, x_i^2, \ldots, x_i^n\} \in \mathbb{R}^n$. Этим объектам соответствует множество классов $\mathbb{Y} = \{y_1, y_2, \ldots, y_L\}$, в котором значения классов $y_i \in \{-1, +1\}$. Назовем множество $\mathbb{X}$ *генеральной выборкой*, и будем считать, что среди объектов нет двух одинаковых.

Если для двух объектов $x_i \in \mathbb{X}$ и $x_j \in \mathbb{X}$ выполняется условие $\forall k = 1, \ldots, n : x_i^k \geq x_j^k$, то будем считать, что $x_i \geq x_j$. Если же $\exists k, t : x_i^k > x_j^k, x_i^t < x_j^t$, то будем считать, что объекты x_i и x_j несравнимы и будем обозначать $x_i \| x_j$.

Пусть также задано множество A, элементы которого называются алгоритмами, где каждый алгоритм $a \in A : \mathbb{R}^n \to \{-1, +1\}$. Существует бинарная функция $I : A \times \mathbb{X} \to \{0,1\}$, называемая *индикатором ошибки*. Если $I(a, x) = 1$, то алгоритм $a \in A$ допускает ошибку на объекте x. Пусть ℓ – фиксированное натуральное число, $\ell < L$. Обозначим через $[\mathbb{X}]^\ell$ множество всех ℓ-элементных подмножеств генеральной выборки $\mathbb{X}$.

Задача обучения по прецедентам состоит в выборе на основе некоторой *обучающей* выборки $X \subset [\mathbb{X}]^{\ell}$ некоторого алгоритма $a \in A$, определяющего для каждого объекта генеральной выборки $\mathbb{X}$ его класс. Выборка $\bar{X} = \mathbb{X} \setminus X$ для выбранного на основе обучающей выборки X алгоритма a называется *контрольной*.

Методом обучения называется отображение $\mu : X \to A$, которое ставит в соответствие обучающей выборке $X \subset [\mathbb{X}]^{\ell}$ некоторый алгоритм $a \in A$. Метод обучения μ называется методом *минимизации эмпирического риска* (МЭР), если

$$\mu(X) \in A(X) = \underset{a \in A}{\operatorname{Arg\,min}} \Big(\sum_{x_i \in X} [a(x_i) y_i < 0] \Big),$$

пессимистичным МЭР, если $\mu^{pes}(X) = \underset{a \in A(X)}{\arg\max} \Big(\sum_{x_i \in \bar{X}} [a(x_i) y_i < 0] \Big)$ и *оптимистичным* МЭР, если $\mu^{opt}(X) = \underset{a \in A(X)}{\arg\min} \Big(\sum_{x_i \in \bar{X}} [a(x_i) y_i < 0] \Big)$. Метод обучения называется *взвешенным* МЭР, если он также учитывает веса объектов:

$$\mu(X) \in A(X) = \underset{a \in A}{\operatorname{Arg\,min}} \Big(\sum_{x_i \in X} w_i [a(x_i) y_i < 0] \Big).$$

1.2. Ограничения монотонности. В качестве множества алгоритмов A в работе рассматривается семейство *монотонных алгоритмов классификации*, то есть $a \in A \Leftrightarrow \big(\forall x_1, x_2 \in \mathbb{R}^n : x_1 \geq x_2 \Rightarrow a(x_1) \geq a(x_2) \big)$.

Монотонная зависимость между признаковым описанием объектов и классами является простым и интуитивно понятным свойством, которым обладают зависимости в различных прикладных областях знаний. Например, чем выше масса тела человека, тем выше риск возникновения сердечно-сосудистых заболеваний. Это иллюстрирует монотонно возрастающую зависимость между признаковым описанием объекта и его классом. Также на практике часто встречается монотонно убывающая зависимость. Без ограничения общности, в работе рассматриваются только монотонно возрастающие зависимости.

1.3. Монотонные корректирующие операции. При решении прикладных задач классификации часто возникает ситуация, когда ни один из существующих

алгоритмов в отдельности не решает задачу с достаточным качеством. В таких случаях пытаются учесть сильные стороны каждого отдельного алгоритма за счет построения из них некоторой композиции. Сам алгоритм композиции, определяющий класс объектов на основе ответов каждого из базовых алгоритмов, входящих в композицию, называют *корректирующей операцией*.

Большинство современных алгоритмов классификации, таких как деревья решений, нейронные сети, регрессии различных типов и многие другие, позволяют определять не только класс объекта, но и свою уверенность в выборе этого класса. Поэтому, корректирующая операция обычно строится не в пространстве классов, определенных каждым из базовых алгоритмов, а в пространстве оценок уверенности базовых алгоритмов в отнесении объектов к классу +1. Это позволяет более тонко учитывать ответы каждого из базовых алгоритмов. Принимается, что если все базовые алгоритмы более склонны к отнесению одного объекта к классу +1, чем другого, то и оценка принадлежности к классу +1 с помощью корректирующей операции для первого объекта должна быть выше. То есть корректирующая операция должна удовлетворять ограничениям монотонности, определенные выше. Будем называть такие композиции *монотонными.*

Использование композиций позволяет снять ограничение на вещественнозначность исходного признакового описания, поскольку многие существующие базовые алгоритмы могут обучаться как на числовых, так и на категорийных признаках объектов. Пространство же оценок базовых алгоритмов является вещественнозначным по определению.

1.4. Проблема переобучения. Семейство монотонных алгоритмов включает в себя большое число различных на обучающей выборке алгоритмов. Поэтому, частота ошибок на обучающей выборке у обученного с помощью МЭР монотонного алгоритма, может быть существенно ниже частоты ошибок на априори неизвестных ему объектах контрольной выборки. В этом случае говорят, что имеет место *переобучение* (overfitting), то есть чрезмерная настройка алгоритма на обучающую выборку.

Для решения этой проблемы в работе рассматривается функционал *полного скользящего контроля* (complete cross-validation, CCV), характеризующий среднюю частоту ошибок алгоритма на контрольной выборке $\bar{X}$ при всевозможных способах разбиения генеральной выборки $\mathbb{X}$ на обучающую и контрольную:

$$CCV = Q_k(\mu, \mathbb{X}) = \frac{1}{C_L^\ell} \sum_{X \cup \bar{X} = \mathbb{X}} \nu(\mu(X), \bar{X}) = E_{\mathbb{X}} \nu(\mu(X), \bar{X}).$$

В этой формуле символом $E_{\mathbb{X}}$ обозначена операция усреднения по всевозможным разбиениям генеральной выборки $\mathbb{X}$ на обучающую и контрольную выборки, а $\nu(a, X)$ означает частоту ошибок алгоритма a на выборке X :

$$\nu(a, X) = \frac{n(a, X)}{|X|} = \frac{1}{|X|} \sum_{x \in X} I(a, x).$$

Чтобы алгоритм не был переобучен, необходимо определить такой метод обучения μ, при котором функционал CCV принимал бы как можно меньшее значение. Поскольку формальный расчет CCV на основе определения требует экспоненциального по количеству объектов генеральной выборки числа операций, то первой задачей является получение вычислительно эффективной точной верхней оценки CCV. Второй задачей является разработка метода минимизации полученной верхней оценки CCV для определения оптимального метода обучения.

Глава 2. Оценки полного скользящего контроля для монотонных классификаторов.

2.1. Методы вычисления комбинаторных оценок CCV. В работе предлагается два различных способа расчета верхних оценок CCV.

Первый способ позволяет рассчитать верхнюю и нижнюю оценки взвешенного МЭР за полиномиальное время, если для всех объектов $\mathbb{X}$ их веса $w_i \in \mathbb{Z}$ и каждый объект описывается единственным числовым признаком, то есть $x_i \in \mathbb{R}$.

Для нахождения оценок CCV свяжем с каждым объектом генеральной выборки x_i два множества:

1. *Множество безошибочных выборок* $E^0(i)$:

$$E^0(i) = \{X : x_i \in \bar{X},\ \forall \mu \in \mathrm{M}\ \ I(\mu(X), x_i) = 0\} \subseteq [\mathbb{X}]^\ell.$$

2. *Множество ошибочных выборок* $E^1(i)$:

$$E^1(i) = \{X : x_i \in \bar{X},\ \forall \mu \in \mathrm{M}\ \ I(\mu(X), x_i) = 1\} \subseteq [\mathbb{X}]^\ell.$$

Теорема 2.1. *Справедливы следующие верхняя и нижняя оценки CCV:*

$$\frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=1}^{L}|E^1(i)| \le Q_k(\mu^{opt}, \mathbb{X}) \le CCV \le Q_k(\mu^{pes}, \mathbb{X}) \le 1 - \frac{1}{k}\frac{1}{C_L^\ell}\sum_{i=1}^{L}|E^0(i)|.$$

Теорема 2.1 формулирует важный принцип расчета значения CCV – от полного перебора по всем разбиениям на обучающую и контрольную выборки можно перейти к расчету вклада в значение оценки CCV каждого объекта, с помощью расчета мощности множества безошибочных и ошибочных выборок. Полученные оценки CCV могут не быть достижимыми, однако они всегда меньше 1. Возможность расчета как нижней, так и верхней оценки CCV позволяет судить о среднем значении CCV. В следующем разделе показывается вычислительно эффективная процедура расчета мощности множества ошибочных и безошибочных выборок.

Второй способ позволяет рассчитать верхнюю оценку CCV для семейства монотонных алгоритмов для многомерной генеральной выборки, используя МЭР, однако не учитывает веса объектов. В отличие от первого способа, где вычисляются оценки для худшего случая, второй способ дополнительно учитывает структуру построения самого монотонного классификатора и допускает представление оценки в аналитическом виде, а также обладает существенно меньшей вычислительной сложностью.

2.2. Одномерная выборка. В одномерном случае, монотонные алгоритмы имеют вид $a(x) = sign(x - c_a)$ и определяются положением порога c_a. Два алгоритма, у которых векторы ошибок на генеральной выборке $\mathbb{X}$ совпадают, являются неразличимыми на этой выборке. Будем полагать, что множество A состоит только из тех алгоритмов, у которых векторы ошибок на $\mathbb{X}$ попарно различны. В этом случае $|A| = L + 1$. Пронумеруем алгоритмы множества $A = \{a_1, a_2, \ldots, a_{L+1}\}$ в

соответствии с возрастанием порога и для определенности выберем значение порога c_i для каждого алгоритма a_i по следующему правилу:

$$c_1 = x_1 - 1;$$
$$c_i = (x_{i-1} + x_i)/2, \quad 1 < i < L;$$
$$c_{L+1} = x_L + 1.$$

Для упрощения записи будем идентифицировать объекты, входящие в обучающую выборку X, по их номеру в обучающей выборке в соответствии с их порядком с помощью верхнего индекса. Например, x^1 означает первый объект, входящий в рассматриваемую обучающую выборку X, а взвешенный МЭР имеет вид $\mu(X) = \arg\min_{a \in A}(\sum_{i=1}^{\ell} w^i [a(x^i) y^i < 0])$.

Без ограничения общности будем считать, что $t \in \{0,1,\ldots,\ell\}$ объектов обучающей выборки X с индексами $j_1, j_2, \ldots j_t$ расположены правее объекта x_i и $\ell - t$ объектов с индексами $n_{\ell-t}, n_{\ell-t-1}, \ldots n_1$ расположены левее объекта x_i (рис. 1).

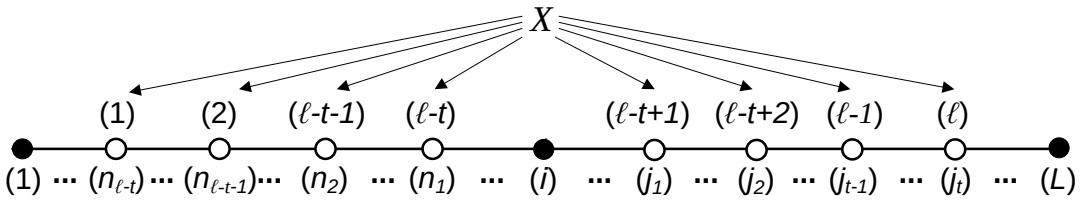


Рис. 1. Расположение объектов обучающей выборки X относительно объекта x_i. Под объектами подписаны их индексы в генеральной выборке. Над объектами – номера в обучающей выборке.

На основе введенных таким образом индексных множеств определим функции, которые будут часто использоваться в дальнейших расчетах:

$$f_+(X) = \min_{k=\ell-t+2,\ldots,\ell+1}\left(\sum_{p=\ell-t+1}^{k-1} y^p w^p\right); \; f_+^0(X) = \min_{k=\ell-t+1,\ldots,\ell+1}\left(\sum_{p=\ell-t+1}^{k-1} y^p w^p\right) = \min(0, f_+(X));$$

$$f_-(X) = \max_{k=1,\ldots,\ell-t}\left(\sum_{p=k}^{\ell-t} w^p y^p\right); \; f_-^0(X) = \max_{k=1,\ldots,\ell-t+1}\left(\sum_{p=k}^{\ell-t} w^p y^p\right) = \max(0, f_-(X));$$

Используя введенные обозначения, доказываются теоремы, определяющие необходимые и достаточные условия, которым должна удовлетворять обучающая выборка для того, чтобы быть безошибочной или ошибочной для объекта x_i.

Теорема 2.2.1. *Обучающая выборка X, в которой t объектов лежат правее объекта x_i, будет являться безошибочной для этого объекта, тогда и только тогда, когда одновременно выполняются следующие условия:*

$$E^0(i)=\left\{X:\begin{array}{l} x_i\in\bar{X}\\ t\in[\min(\ell,L-i),\max(\ell-i+1,0)]\\ f_-(X)+f_+^0(X)>0,\text{ если } y_i=+1\\ f_-^0(X)+f_+(X)<0,\text{ если } y_i=-1\end{array}\right\}$$

Теорема 2.2.2. *Обучающая выборка X, в которой t объектов лежат правее объекта x_i, будет являться ошибочной для этого объекта, тогда и только тогда, когда одновременно выполняются следующие условия:*

$$E^1(i)=\left\{X:\begin{array}{l} x_i\in\bar{X}\\ t\in[\min(\ell,L-i),\max(\ell-i+1,0)]\\ f_-^0(X)+f_+(X)<0,\text{ если } y_i=+1\\ f_-(X)+f_+^0(X)>0,\text{ если } y_i=-1\end{array}\right\}$$

Для упрощения формулы расчета мощности множеств ошибочных и безошибочных выборок обозначим операцию суммирования по всевозможным выборкам $X^+\subset[\mathbb{X}]^t$, лежащим правее объекта x_i, символом $\sum_i^+$. Символом $\sum_i^-$ обозначим операцию суммирования по всевозможным выборкам $X^-\subset[\mathbb{X}]^{\ell-t}$, лежащим левее объекта x_i. В этих обозначениях введем функции:

$$\mathrm{f}_+^0(i,t,s)=\sum\nolimits_i^+[f_+^0(X^+)=s];\ \mathrm{f}_+(i,t,s)=\sum\nolimits_i^+[f_+(X^+)=s];$$
$$\mathrm{f}_-^0(i,t,s)=\sum\nolimits_i^-[f_-^0(X^-)=s];\ \mathrm{f}_-(i,t,s)=\sum\nolimits_i^-[f_-(X^-)=s].$$

Мощность множеств ошибочных и безошибочных выборок для объекта x_i рассчитывается на основе следующих формул:

$$P=\sum_{t=\min(\ell,L-i)}^{\max(\ell-i+1,0)}\left([t=0]\sum_{s>0}\mathrm{f}_-(i,\ell,s)+[0<t<\ell]\sum_{s\in S}\left(\mathrm{f}_+^0(i,t,s)\sum_{s'>-s}\mathrm{f}_-(i,\ell-t,s')\right)\right)$$
$$N=\sum_{t=\min(\ell,L-i)}^{\max(\ell-i+1,0)}\left([0<t<\ell]\sum_{s\in S}\left(\mathrm{f}_-^0(i,\ell-t,s)\sum_{s'<-s}\mathrm{f}_+(i,t,s')\right)+[t=\ell]\sum_{s<0}\mathrm{f}_+(i,\ell,s)\right)$$

$$|E^0(i)|=\begin{cases}P, y_i=+1\\ N, y_i=-1\end{cases}; \qquad |E^1(i)|=\begin{cases}N, y_i=+1\\ P, y_i=-1\end{cases}.$$

Значения функций f рассчитываются рекурсивно на основе следующих теорем:

Теорема 2.2.3. $\mathrm{f}_+(i,t,s)=\mathrm{f}_+^0(i+1,t-1,s-w_{i+1}y_{i+1})+\mathrm{f}_+(i+1,t,s)$.

Теорема 2.2.4. $\mathrm{f}_-(i,t,s)=\mathrm{f}_-^0(i-1,t-1,s-w_{i-1}y_{i-1})+\mathrm{f}_-(i-1,t,s)$.

Вычислительная сложность расчета оценок CCV на основе приведенных формул расчета мощности ошибочных и безошибочных множеств равна $O(L\ell\sum_{i=1}^{L}w_i)$. При этом затраты памяти, необходимой для расчета, имеют такую же оценку $O(L\ell\sum_{i=1}^{L}w_i)$.

Если брать веса объектов из множества $w\in\{0,1\}$, то есть принимать решение об использовании того или иного объекта для обучения, то вычислительная сложность процедуры вычисления CCV будет порядка $O(\ell L^2)$, а затраты памяти – $O(\ell L^2)$.

Основываясь на этой процедуре, был предложен и реализован метод жадного спуска для подбора оптимального набора весов объектов из множества $w\in\{0,1\}$, минимизирующего значение верхней оценки CCV. Результаты экспериментов на модельных задачах показали, что предложенный метод действительно позволяет фильтровать шумовые объекты, тем самым уменьшая риск переобучения.

2.3. Многомерная выборка. Аналогично одномерному случаю будем считать, что все монотонные алгоритмы из множества A различимы на генеральной выборке $\mathbb{X}\subset\mathbb{R}^n$, то есть их векторы ошибок на выборке $\mathbb{X}$ попарно различны. В этом случае любой монотонный алгоритм $a\in A$ полностью определяется двумя непересекающимися множествами:

$$\Omega_+=\{x\in\mathbb{X}: a(x)=+1\};$$

$$\Omega_-=\{x\in\mathbb{X}: a(x)=-1\}.$$

Эти множества должны обладать свойством, необходимым и достаточным для монотонности алгоритма a: $\forall x_1 \in \Omega_-, \forall x_2 \in \Omega_+ : x_2 > x_1 \vee x_2 \| x_1$.

Тогда задача метода обучения μ, минимизирующего эмпирический риск, состоит в построении таких множеств Ω_- и Ω_+, обладающих описанным выше свойством, для которых число ошибок монотонного классификатора минимально.

Назовем пару индексов (i, j) *дефектной*, если выполняется одно из условий $x_i > x_j, y_i < y_j$ или $x_i < x_j, y_i > y_j$.

Теорема 2.3.1. *Вычислительная сложность обучения монотонного алгоритма a, минимизирующего эмпирический риск, имеет порядок $O(m\sqrt{d})$, где m – число дефектных пар, образованных объектами генеральной выборки $\mathbb{X}$, а d – число объектов генеральной выборки, образующих дефект.*

Доказательство этой теоремы основано на конструктивном методе нахождения оптимального монотонного алгоритма с помощью сведения исходной задачи к задаче поиска минимального вершинного покрытия в графе, образованном всеми дефектными парами исходной генеральной выборки. Поскольку этот граф является двудольным, то для решения этой задачи существует эффективный алгоритм Хопкрофта-Карпа. В худшем случае, вычислительная сложность нахождения оптимального монотонного алгоритма с помощью Теоремы 2.3.1 есть $O(d^2\sqrt{d})$. Предложенный метод находит монотонный алгоритм с помощью МЭР быстрее наилучшего из существовавших ранее методов, сложность которого оценивается как $O(d^3)$.

Назовем объект x_i *дефектным*, если он входит хотя бы в одну дефектную пару и *бездефектным*, если он не входит ни в одну дефектную пару. Обозначим D_0 – множество всех дефектных объектов генеральной выборки.

Назовем *клином* объекта $x_i \in \mathbb{X}$ множество:

$$W(x_i) = \begin{cases} x_k \in \mathbb{X} \mid x_i < x_k ; y_i = y_k = -1; \\ x_k \in \mathbb{X} \mid x_k < x_i ; y_i = y_k = +1. \end{cases}$$

Бездефектным клином объекта $x_i \in \mathbb{X}$ назовем множество $\overline{W}(x_i) = W(x_i)/D_0$. Обозначим w_i мощность клина объекта x_i, а $\overline{w}_i$ – мощность бездефектного клина объекта x_i.

Для устранения неопределенности в выборе оптимального алгоритма с помощью МЭР, рассмотрим подсемейство монотонных классификаторов *ближайшего соседа*, имеющих вид:

$$a(x) = y(\arg\min_{x_j \in U} \rho(x, x_j)),$$

где U – некоторая монотонная подвыборка $\mathbb{X}$, функция расстояния от классифицируемого объекта x до объекта $x_j \in U$ зависит от класса объекта x_j и определяется следующим образом:

$$\rho(x, x_j) = \begin{cases} \max(x_j^1 - x^1, \ldots, x_j^n - x^n, 0), & y(x_j) = +1 \\ \max(x^1 - x_j^1, \ldots, x^n - x_j^n, 0), & y(x_j) = -1 \end{cases}.$$

В работах К.В. Воронцова было доказано, что построенный таким образом алгоритм $a(x)$ будет монотонным на $\mathbb{R}^n$, если он является монотонным на всех объектах $\mathbb{X}$.

Множество объектов $\mathbb{X}$, на которых алгоритм $\mu(\mathbb{X})$, минимизирующий эмпирический риск, ошибается, называется *множеством немонотонности*. Обозначим его D. *Степень немонотонности* δ генеральной выборки $\mathbb{X}$ определяется как $|D|/L$.

Заметим, что выборка $\mathbb{X}/D$ является монотонной, то есть $\forall x_1, x_2 \in \mathbb{X}/D$: $x_1 \geq x_2 \Rightarrow y_1 \geq y_2$. Для каждого объекта $x_i \in \mathbb{X}$ удалим из генеральной выборки множество $D/\{x_i\}$. Обозначим D_i мощность множества $D/\{x_i\}$, то есть $D_i = |D| - [x_i \in D]$. Оставшиеся объекты упорядочим по возрастанию расстояния от объекта x_i, пронумеровав их двойными индексами: $x_{i,1}, \ldots, x_{i,L-1-D_i}$. Таким образом,

$$\rho(x_i, x_{i,1}) \leq \rho(x_i, x_{i,2}) \leq \ldots \leq \rho(x_i, x_{i,L-1-D_i}).$$

Обозначим $I_m(x_i)$ ошибку, возникающую, если правильный ответ y_i заменить ответом на его m-ом соседе :

$$I_m(x_i) = [y_i \neq y_{i,m}] \quad i = 1,\ldots,L \quad m = 1,\ldots,L-1-D_i$$

Используя введенные обозначения, доказывается верхняя оценка CCV.

Теорема 2.3.2. *Если метод* μ *минимизирует эмпирический риск в классе всех монотонных алгоритмов, то:*

$$CCV \leq \sum_{i=1}^{L} \frac{1}{L} \left(\sum_{d=\max\{1,\, D_i+\bar{w}_i+1-k\}}^{D_i} \frac{C_{D_i}^{d} C_{L-1-D_i-\bar{w}_i}^{\ell-d}}{C_{L-1}^{\ell}} + [D_i + \bar{w}_i < k] \sum_{m=1}^{k-D_i} \frac{I_m(x_i) C_{L-1-D_i-m}^{\ell-1}}{C_{L-1}^{\ell}} \right).$$

Доказанная оценка состоит из двух слагаемых. Первое слагаемое объясняет зависимость CCV от степени немонотонности генеральной выборки. Чем больше мощность множества немонотонности D, тем больший вклад вносит это слагаемое в оценку CCV. Поэтому, назовем первое слагаемое *немонотонной* частью оценки CCV. Второе слагаемое объясняет зависимость CCV от структуры монотонного алгоритма. Его значение тем больше, чем больше объектов оказывается рядом в смысле введенного с помощью функции ρ расстояния. Поэтому, назовем второе слагаемое *некомпактной* частью оценки CCV. Поскольку оценка CCV учитывает как свойства выборки, так и структуру алгоритмов, то назовем ее *гибридной*. Основное преимущество гибридной оценки в том, что чем более монотонная выборка, тем ближе эта оценка к точному значению CCV. Если генеральная выборка является монотонной, то есть $|D| = 0$, то гибридная оценка совпадает с точным значением CCV.

Проведенные численные эксперименты на модельных задачах показали, что точность гибридной оценки CCV превосходит точность другой комбинаторной оценки CCV, полученной К. В. Воронцовым, если размерность признакового описания объектов больше двух. Причем чем больше размерность признакового описания объектов, тем гибридная оценка CCV становится более точной.

Глава 3. Методы построения монотонных композиций классификаторов.

3.1 Монотонная коррекция при независимом обучении базовых алгоритмов является самым простым способом построения монотонной композиции. Все

базовые алгоритмы обучаются независимо на обучающей выборке, и в пространстве их оценок принадлежности объектов к классу +1 на основе Теоремы 2.2.1 строится монотонный классификатор ближайшего соседа.

Эксперименты по оценке качества монотонной композиции проводились на реальных задачах из репозитария UCI. Качество монотонной композиции сравнивалось с двумя другими классическими композициями: взвешенным голосованием оценками принадлежности к каждому классу и выбором класса с максимальной оценкой.

В качестве базовых алгоритмов использовались следующие классические алгоритмы: решающие деревья C50, CART, QUEST, CHAID; нейронная сеть на основе многослойного персептрона; к-ближайших соседей, логистическая регрессия и SVM с автоматическим выбором функции ядра.

Эксперименты показали, что даже простейшая монотонная композиция способна уменьшать частоту ошибок на контрольной выборке по сравнению с базовыми алгоритмами и другими композициями. Однако, на всех исследуемых задачах средняя частота ошибок на контроле оказалась в разы больше средней частоты ошибок на обучении, то есть композиция переобучалась.

3.2. Монотонный бустинг. Для уменьшения риска переобучения предлагается три подхода к построению монотонных композиций, в которых обучение базовых алгоритмов не является независимым.

Первый подход основан на идее минимизации количества дефектных пар в пространстве оценок базовых алгоритмов. Для этого базовые алгоритмы настраиваются последовательно, и каждый последующий базовый алгоритм настраивается на объекте обучающей выборке тем сильнее, чем в большее число дефектных пар входит этот объект. Преимуществом построения такой композиции является простота ее реализации. Однако проведенные вычислительные эксперименты показали, что при количестве базовых алгоритмов больше двух, построенная таким способом композиция начинает сильно переобучаться.

Для уменьшения риска переобучения предлагается второй подход, основанный на минимизации комбинаторной оценки CCV, доказанной в работах К. В.

Воронцова [X], при последовательном добавлении базовых алгоритмов в композицию. Недостатком этого подхода является относительно низкое качество классификации построенной с его помощью монотонной композиции.

Третий подход основан на идее построения базовых алгоритмов таким образом, чтобы в пространстве оценок гибридная оценка CCV монотонного классификатора, доказанная в Теореме 2.3.2, была минимальна. Для этого построение всех базовых алгоритмов происходит одновременно, причем каждый базовый алгоритм настраивается на объектах генеральной выборки с уникальными для него весами. Поэтому, параметрами МЭР является матрица $L \times m$ весов объектов обучающей выборки, где m – количество алгоритмов в композиции.

Метод минимизации гибридной оценки CCV основан на принципах метода градиентного спуска, реализованных для дискретного случая. Изначально вся матрица весов объектов инициализируются произвольными значениями из отрезка [1,2]. Аналитически оценивается вклад каждого объекта генеральной выборки в величину текущего значения гибридной оценки CCV. Эвристически оценивается направление изменения весов для каждого объекта, которое привело бы к максимальному уменьшению гибридной оценки CCV, если бы веса всех остальных объектов были бы фиксированы. Используя величину вклада объекта в гибридную оценку CCV в качестве модуля вектора изменения и рассчитанное направление изменения, проводится итеративный пересчет весов. Процедура останавливается, когда не удается улучшить гибридную оценку CCV больше, чем заданное число итераций подряд.

Каждый из подходов был реализован на языке R, и качество его классификации было проверено на модельных задачах различной степени трудности, а также реальных задачах из различных прикладных областей. В качестве базовых алгоритмов использовался алгоритм логистической регрессии.

Результаты экспериментов показали, что монотонный бустинг, построенный с помощью третьего подхода, превосходит по качеству как алгоритмы, построенные с помощью первых двух подходов, так и алгоритм бустинга AdaBoost. Он требует существенно меньшего числа базовых алгоритмов в композиции и при

числе базовых алгоритмов, большем трех, риск переобучения для него практически отсутствует.

В заключении изложены основные результаты диссертации.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Получена комбинаторная оценка полного скользящего контроля для семейства монотонных алгоритмов, учитывающая как свойства задачи, так и свойства самого семейства монотонных алгоритмов, и проведены эксперименты, доказывающие точность этой оценки.
2. Предложен вычислительно эффективный метод построения оптимального монотонного классификатора, минимизирующего эмпирический риск.
3. Предложен метод построения монотонной композиции базовых алгоритмов, минимизирующий полученную оценку полного скользящего контроля для уменьшения переобучения и повышения качества всей композиции.
4. С помощью вычислительных экспериментов показано, что предложенный метод построения монотонной композиции повышает качество классификации по сравнению с отдельными базовыми алгоритмами, а также некоторыми другими известными методами построения композиций.

ПУБЛИКАЦИИ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

1) Гуз И. С. Гибридные оценки полного скользящего контроля для монотонных классификаторов. // Докл. всеросс. конф. Математические методы распознавания образов-15. — М.: МАКС Пресс, 2011. — С. 98–103.

2) Гуз И. С. Минимизация эмпирического риска при построении монотонных композиций классификаторов // **Труды МФТИ** –2011.– Т.3, №3 (11) – С. 115-121.

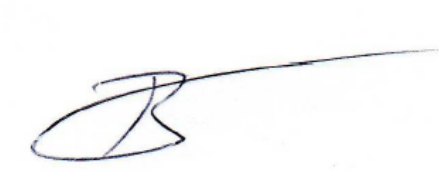
3) Гуз И. С. Конструктивные оценки полного скользящего контроля для пороговой классификации // **Математическая биология и биоинформатика**. — Т.6. — В.2. — 2011. — http://www.matbio.org/2011/Guz2011(6_173).pdf

4) Гуз И.С. Исследование обобщающей способности семейства монотонных функций // Моделирование и обработка информации: Сб.ст./Моск.физ.-тех. ин-т. – М., 2008. – С. 114-120.

5) Гуз И. С. Обобщающая способность монотонных композиций классификаторов // Докл. межд. конф. Интеллектуализация обработки информации, ИОИ-7, — Симферополь. 2008, — С. 75-77.

6) Гуз И. С. Нелинейные монотонные композиции классификаторов // Докл. всеросс. конф. Математические методы распознавания образов-13. — М.: МАКС Пресс, 2007. — С. 111–114.

7) Гуз И. С. Алгоритмические композиции с монотонными и выпуклыми корректирующими операциями //Современные проблемы фундаментальных и прикладных наук. Часть VII. Прикладная математика и экономика: Труды XLV научной конференции. /Моск. физ. – техн. ин-т. – М. – Долгопрудный, 2006. – С. 282-283.



ГУЗ Иван Сергеевич

Комбинаторные оценки обобщающей способности и методы обучения монотонных классификаторов

Автореферат

Подписано в печать 31.10.2011 г. Формат 60x90 1/16.
Усл. печ. л. 1,0. Тираж 100 экз. Заказ № 524.
Отпечатано в типографии «Реглет»
119526, г. Москва, Страстной бульвар, д.6,стр. 1
(495) 978-43-34; www.reglet.ru