ОТБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ: ПОСТАНОВКА ЗАДАЧИ И МЕТОДИКА ЕЕ РЕШЕНИЯ

О. В. Жвалевский

Санкт-Петербургский институт информатики и автоматизации РАН

СПИИРАН, 14-я линия ВО, д. 39, Санкт-Петербург, 199178

<ozh@spiiras.nw.ru>

УДК 519.237.8+519.25

Жвалевский О. В. **Отбор информативных признаков: постановка задачи и методика ее решения** // Труды СПИИРАН. 2007. Вып. 4. — СПб.: Наука, 2007.

Аннотация. *В работе рассматривается задача отбора информативных признаков из числа тех, которые вычисляются при обработке тензометрических данных методом анализа фрактальной динамики. Обсуждаются способы вычисления признаков, методы классификации и критерии отбора признаков. Приводятся основные алгоритмы отбора информативных признаков.* — Библ. 7 назв.

UDC 519.237.8+519.25

Zhvalevsky O. V. **Features extraction: problem formulation and solving technique** // SPIIRAS Proceedings. 2007. Issue 4. — SPb.: Nauka, 2007.

Abstract. *The application of the fractal dynamic analysis method to tensometrical data is results in some variables appearance. There is a problem of features extraction in the paper. Features calculation methods, classification methods and features extraction criteria are discussed. The review of features extraction algorithms are represented.* — Bibl. 7 items.

1. Введение

При обработке тензометрических данных [1] методом анализа фрактальной динамики [2] вычисляются признаки, которые в своем исходном виде не позволяют удовлетворительно решать задачу классификации испытуемых [3] для целей диагностики болезни Паркинсона. Возникает необходимость в разработке процедуры отбора таких сочетаний признаков, для которых количество ошибок классификации будет минимальным. Найденные сочетания признаков будем называть *оптимальными наборами*, а признаки, входящие в оптимальные наборы, мы будем называть *информативными*.

При вычислении признаков по методу анализа фрактальной динамики (АФД) исследуются:

1) структура тензометрических данных — разложение анализируемого сигнала на *компоненты* (раздел **2**);
2) конструктивные параметры в алгоритме метода АФД — *двухпараметрическая модель* огибающей спектра мощности, *сегментация* процесса и *собственные числа* процесса (раздел **3**);
3) способы обращения с объектами различных классов — *бинарная* и *множественная* классификации (раздел **4**);
4) способы классификации объектов — *линейные* и *нелинейные* классификаторы (раздел **5**);
5) критерии отбора информативных признаков — *внутренние* и *внешние* критерии отбора (раздел **6**).

Для каждого способа вычисления признаков, для каждого классификатора и для каждого критерия отбора может быть найден свой оптимальный набор. По-

этому следует указывать условия, при которых данный оптимальный набор был получен.

В работе формулируется задача отбора информативных признаков (раздел **7**) как задача поиска наилучших условий и предлагается методика ее решения (раздел **8**) посредством применения ряда алгоритмов (раздел **9**).

2. Тензометрические данные

Тензометрические данные — это данные, полученные при регистрации пьезорезистивным датчиком усилия, удерживаемого испытуемым в ходе специальным образом поставленного измерительного эксперимента или *теста*. *Треморограмма* — это результат отдельного теста. Существуют четыре стандартных теста, представленные в табл. 1.

Таблица 1

Четыре стандартных теста (типы измерительных экспериментов)

№	Тест	Удерживаемое усилие
1	для пальцев	минимально
2	для пальцев	максимально
3	для вытянутых рук	минимально
4	для вытянутых рук	максимально

Каждая треморограмма представляет собой запись (см. табл. 2), полученную на выходе четырехканального аналого-цифрового преобразователя (АЦП) и содержащую как *произвольный компонент усилия* (каналы 3–4), уровень которого и задаётся экспериментатором, так и *непроизвольный компонент усилия* (каналы 1–2), который содержит, собственно, *тремор*.

Таблица 2

Каналы АЦП: регистрируемые компоненты

№	Канал	Компонент
1	тремор левой руки	непроизвольный
2	тремор правой руки	
3	усилие левой руки	произвольный
4	усилие правой руки	

При вычислении признаков по методу АФД, следовательно, необходимо оценить информативность как произвольного, так и непроизвольного компонентов усилия. Поэтому возникает по крайней мере три варианта вычисления признаков (табл. 3).

Таблица 3

Три стандартных варианта вычисляемых признаков,
где M — число признаков

№	Компонент(ы)	Каналы	M
1	только тремор	1–2	16
2	только усилие	3–4	16
3	и тремор, и усилие	1–4	22

Следовательно, в нашем распоряжении оказываются 3 набора признаков: 2 набора, состоящих из $M = 16$ признаков, и 1 набор, состоящий из $M = 22$ признаков. На рис. 1 в компактной форме показаны все вычисляемые по методу

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

знаков. На рис. 1 в компактной форме показаны все вычисляемые по методу АФД признаки с учетом всех трех вариантов.

Признаки	**порядковые номера признаков**
модель огибающей спектра мощности	[1 2 3 4 5 6]
остатки	[7 8 9]
сингулярные числа	[10 11 12 13] ([10 11 12 13 14 15 16 17])
отношения собственных чисел	[14 15] ([18 19 20 21])
разброс	[16] ([22])

Рис. 1. Признаки, вычисляемые по методу анализа фрактальной динамики (АФД) [3]. Порядковые номера (справа) соответствуют порядку признаков в матрицах данных.

3. Сегментация процесса

В методе АФД производится *сегментация* — разбиение анализируемого временного ряда на интервалы или *сегменты*. На каждом сегменте вычисляются свои признаки, и только после осреднения полученных признаков вычисляются общие признаки для анализируемого временного ряда в целом. Длина сегмента выбирается так, чтобы процесс на каждом сегменте был квазистационарным. Это основное требование при применении метода АФД. В алгоритме метода АФД используется *односекундная* сегментация. В нашем случае разбиение на односекундные интервалы является некоторым произволом и не опирается на исследование стохастических свойств процесса.

Параметры сегментации по-разному влияют на вычисляемые признаки: *число отсчетов* на сегменте влияет на достоверность спектральных характеристик (чем больше отсчетов, тем лучше), в то время как *число сегментов* влияет на осреднение (чем больше сегментов, тем лучше). Поэтому окончательный выбор сегментации — результат компромисса, основанного на подробном исследовании. Строго говоря, выбор длины сегмента — отдельная исследовательская задача, которую следует решать, например, методами обобщенного спектрального анализа, *независимо* от методов распознавания образов. Однако возможен и другой подход: последовательный перебор нескольких *произвольных* вариантов разбиения анализируемых временных рядов и рассмотрение признаков, вычисляемых для каждого разбиения.

Будем использовать три сегментации, представленные в табл. 4. Одна из них — стандартная, односекундная, а две другие — дополнительные.

Таблица 4

Три сегментации для стандартных треморограмм

№	**Длина сегмента в секундах (с)**	**Число отсчетов на сегменте**	**Количество сегментов**
1	1	100	30
2	2	200	15
3	0.5	50	60

Таким образом, необходимо исследовать вопрос о том, какая из предложенных сегментаций является наиболее предпочтительной при вычислении признаков.

4. Шкала наименований

При диагностике болезни Паркинсона шкала наименований состоит из четырех наименований, представленных в табл. 5.

Таблица 5

Классы испытуемых (шкала наименований)

№	Класс
1	Здоровые испытуемые
2	Больные испытуемые
3	Испытуемые, больные болезнью Паркинсона
4	Испытуемые, больные синдромом паркинсонизма

Строго говоря, классификатор должен относить предъявленные ему объекты к одному из четырех классов. Это означает, что обучающие выборки должны состоять из объектов всех четырех классов. Классификацию с числом классов большим 2 будем здесь называть *множественной*. Проведение множественной классификации необходимо для более тонкой классификации по сравнению с классификацией объектов двух классов, но при этом, однако, затрудняется построение разделяющих поверхностей, особенно в случае применения линейного классификатора. Поэтому, выбирая *линейный* классификатор, приходится ограничиваться только классификацией объектов каких-либо *двух* представляющих интерес классов. Например, классификации подвергаются объекты только классов 1 и 3 или только классов 3 и 4 (1 и 4). Также возможно объединение объектов нескольких исходных классов — классов 1–4 — в один *суперкласс* и проведение классификации объектов, принадлежащих сформированным суперклассам. Следовательно, задача множественной классификации может быть сведена к задаче классификации объектов двух классов методом иерархической классификации с образованием суперклассов и последовательной дихотомией объектов, относящихся к различным исходным классам.

В контексте диагностики БП следует также учитывать и ту ошибку, которую вносит в процесс классификации применение лекарственных препаратов во время проведения измерительного эксперимента. Это заведомо снижает количество правильно распознанных объектов, поскольку, с одной стороны, учитываются ошибочные сведения обучающей выборки и, с другой стороны, принимается ошибочное решение об отнесении некоторых объектов контрольной выборки. Данная проблема частично преодолевается разбиением классов 2–4 на класс объектов «без применения лекарственных препаратов» («фон») и на класс объектов «с применением лекарственных препаратов (с указанием конкретного препарата)». Это приводит к сокращению генеральных совокупностей, поэтому необходимо ставить и проводить новые измерительные эксперименты, что оказывается за пределами настоящего исследования.

5. Методы классификации

Задача классификации решается методами статистической теории распознавания образов [4]. При этом наследуются основные ограничения, свойст-

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

венные статистическим методам, а получаемые результаты далеко не всегда обладают требуемой точностью и достоверностью. Это вызвано, с одной стороны, тем, что большинство статистических выводов делается для нормально распределенных случайных величин, причем наименее разработанными оказываются методы многомерного анализа, хотя именно многомерный (существенно) случай и представляет интерес на практике. С другой стороны, это вызвано тем, что объемы генеральных совокупностей, с которыми иногда приходится иметь дело (а это, как правило, десятки или даже сотни объектов), оказываются недостаточными для получения точных и достоверных статистических выводов.

При использовании стандартной функции MATLAB `classify` мы фактически имеем дело с тремя различными *линейными* классификаторами (см. табл. 6). Поэтому каждая матрица данных может быть обработана тремя различными способами.

Таблица 6

Три стандартных классификатора

№	Ккассификатор
1	`linear`
2	`quadratic`
3	`mahalanobis`

Стандартные классификаторы MATLAB отличаются друг от друга способами вычисления расстояния между объектами и способами построения ковариационных матриц, поэтому они заведомо должны дать (что и подтверждается на практике) различные результаты классификации на одних и тех же объектах.

Таким образом, необходимо исследовать, какой из стандартных классификаторов дает наилучшие результаты при распознавании.

При полном исследовании необходимо проводить предварительный анализ матриц данных, что потребует привлечения методов теории кластеризации.

6. Критерии отбора

Отбор информативных признаков производится в соответствии с некоторым критерием. Этот критерий формулируется так, чтобы искомое решение доставляло оптимум (минимум или максимум) соответствующего функционала. Количество ошибок при классификации объектов выборки — это один из самых простых функционалов, а критерий (основанный на нем) — самый простой критерий отбора. Вместо числа ошибок можно рассматривать процент правильно распознанных объектов, причём, если речь идет о классификации объектов двух и более классов, следует учитывать наихудший процент правильно распознанных объектов из числа тех, которые получены для всех классов.

Минимум ошибок классификации на обучающей выборке — это *внутренний критерий* отбора, в то время как минимум ошибок классификации на контрольной выборке — это *внешний критерий отбора* [5]. Внутренний критерий позволяет отсеять заведомо неоптимальные наборы. Задавая некоторую нижнюю границу процента правильно распознанных объектов, мы сужаем число тех наборов, среди которых производится поиск оптимального. Каждый такой набор затем проверяется в соответствии с внешним критерием.

В ходе предварительного рассмотрения мы ограничены лишь самыми простыми критериями, однако при полном исследовании необходимо рассмат-

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

ривать более сложные критерии, основанные на всевозможных разбиениях генеральных совокупностей на обучающие и контрольные выборки и методиках осреднения результатов, полученных на элементах разбиения. Для этого необходимо иметь запас контрольных объектов. При этом акцент смещается на методики манипулирования (малыми) генеральными совокупностями.

7. Постановка задачи

Таким образом, у нас есть три варианта вычисления признаков, связанных со структурой исходных данных (треморограмм), три варианта вычисления признаков, связанных с выбором из три заданных сегментаций. Это даёт 9 вариантов признаков.

Три стандартных классификатора приводят к появлению трех вариантов обработки каждой матрицы данных, а значит, и к различным оптимальным наборам. Всего имеется, таким образом, 27 *различных* вариантов обработки, которые приходится рассматривать *раздельно*.

Таким образом, ставится задача: найти оптимальный набор признаков и указать наилучшие условия, при которых он был получен. Следовательно, в результате исследования, необходимо выбрать:

1) наилучший способ вычисления признаков;
2) наилучший способ классификации.

Оптимальность понимается в смысле минимума количества ошибок классификации. Именно этот критерий качества классификации и является определяющим при выборе 1) и 2).

Если привлекать и другие классификаторы, исследовать влияние на качество классификации множества конструктивных параметров в алгоритме метода АФД, то число вариантов станет необозримым, а сам поиск вариантов — технически неосуществимым. Поэтому приходится ограничиваться лишь самыми простыми критериями отбора.

Процедура отбора информативных признаков также должна быть оптимальной. Однако здесь оптимальность понимается в ином смысле: алгоритм поиска должен решать задачу за обозримое время, а само доставляемое им решение должно быть *эффективным*, то есть близким к оптимальному.

8. Методика решения

Универсальный способ решения задачи — это *полный перебор* вариантов. В тех случаях, когда полный перебор невозможен, прибегают к эвристическим процедурам, существенно сокращающим объем вычислений. Таковы алгоритмы *последовательного перебора* вариантов, основанные на включении и/или исключении признаков. *Случайный поиск с адаптацией* представляет собой некоторую альтернативу перечисленным методам, обладая вычислительной эффективностью и близостью результатов к оптимальным.

Полный перебор вариантов — тривиальный и результативный способ отбора признаков. Он, очевидно, пригоден для малого числа признаков, поэтому для $M = 16$ или $M = 22$ он оказывается неосуществим за обозримое время. Тем не менее, будучи осуществлен, такой перебор может дать представление о сравнительной ценности различных наборов.

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Вместо того чтобы перебирать все варианты, можно начать с некоторого фиксированного набора признаков и последовательными операциями дополнения и/или исключения признаков добиться улучшения качества классификации. В ходе выполнения того или иного алгоритма последовательного перебора вариантов в общем случае получаются неоптимальные наборы, которые могут существенно отличаться от оптимальных по составу признаков. Тем не менее время поиска оказывается существенно меньшим. Таким образом, последовательный перебор — менее результативный, чем полный перебор, но технически осуществимый (для большинства случаев) способ отбора признаков, основанный на включении и/или исключении признаков.

Случайный поиск с адаптацией (СПА) [6] состоит в том, что признакам сопоставляются некоторые веса, которые определяют степень их информативности. В противоположность методам последовательного перебора в методе СПА всегда используется одно и то же (заранее заданное) количество признаков, что особенно важно для сохранения соотношения между числом признаков и числом объектов в обучающей выборке. (Другими словами, размерность признакового пространства остается неизменной.) В противоположность методу полного перебора метод СПА позволяет существенно сократить объем вычислений, что бывает особенно заметно при числе признаков в искомом наборе, вдвое меньшем, чем число признаков в исходном наборе. Поэтому СПА — это эффективный в вычислительном отношении способ отбора информативных признаков, который позволяет в ряде случаев получать оптимальные наборы.

9. Алгоритмы поиска оптимальных наборов

Основные алгоритмы поиска оптимальных наборов перечислены в табл. 7. Изложение алгоритмов следует, в основном, обзору, данному в [7], частично опираясь на сведения, содержащиеся в [5].

Таблица 7

Алгоритмы отбора информативных признаков:
стандартные (станд.) и модифицированные (модиф.)

№	Алгоритм	Станд.	Модиф.
1	полного перебора вариантов	`Full`	
2	последовательного добавления признаков	`Add`	`Add*`
3	последовательного исключения признаков	`Del`	`Del*`
4	`Add` с последующим исключением	`AddDel`	`AddDel*`
5	`Del` с последующим добавлением	`DelAdd`	`DelAdd*`
6	случайный поиск с адаптацией	`Prob`	

Разберем основные алгоритмы подробнее.

9.1. Алгоритм `Full`

Алгоритм `Full` последовательно просматривает наборы, состоящие ровно из одного элемента. Затем берутся двухэлементные наборы, трехэлементные наборы и т. д. В итоге, задача поиска оптимального набора разбивается на M подзадач, в каждой из которых ищутся оптимальные наборы, состоящие ровно из m признаков, где $m \in \overline{1,M}$. Результатом работы алгоритма будет один или несколько наборов, дающих наилучшее качество классификации в смысле вы-

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

бранного критерия качества классификации. Если нас интересует набор, состоящий ровно из k признаков (для некоторого $k \in \overline{1,M}$), то поиск ограничивается C_M^k сочетаниями признаков из M по k.

Рассмотрим некоторые результаты (рис. 2) вычислений, полученные для выборки, состоящей из 60 объектов (30 здоровых испытуемых и 30 испытуемых, больных болезнью Паркинсона) и представляют собой результаты «на обучении». Слева представлены «малые», одноэлементные наборы. Выделены подчеркиванием наилучшие результаты классификации, в то время как наихудшие результаты выделены жирным шрифтом. Справа — 15-элементные наборы: набор «1⁻» соответствует набору, состоящему из 15 признаков, который получается из полного набора из 16 признаков вычеркиванием признака 1.

п	1	3	1	3	1	3	п	1	3	1	3	1	3
1	36	53	36	86	96	10	1⁻	86	80	—	—	96	96
2	43	63	26	80	63	46	2⁻	80	80	—	—	90	100
3	33	70	26	90	96	23	3⁻	86	83	—	—	100	100
4	26	96	26	96	30	90	4⁻	86	80	—	—	100	100
5	60	60	66	46	60	60	5⁻	—	—	—	—	86	100
6	16	100	26	96	30	96	6⁻	—	—	—	—	96	96
7	83	33	13	96	96	10	7⁻	86	80	—	—	93	100
8	30	73	10	93	83	20	8⁻	—	—	—	—	93	93
9	**16**	**96**	16	96	26	96	9⁻	—	—	—	—	93	96
10	53	50	73	36	40	66	10⁻	90	80	—	—	100	100
11	63	43	63	43	63	43	11⁻	—	—	—	—	93	100
12	63	53	43	66	**100**	**3**	12⁻	86	80	—	—	100	93
13	63	53	73	73	90	26	13⁻	—	—	—	—	93	100
14	26	93	90	03	**3**	**100**	14⁻	—	—	—	—	93	93
15	26	83	13	96	26	83	15⁻	—	—	—	—	93	100
16	80	60	86	30	76	70	16⁻	—	—	—	—	96	100

Рис. 2. Признаки п и результаты классификации объектов классов 1 и 3.

Рассмотрим сначала одноэлементные наборы (см. рис. 2, слева).

Очевидно, что сами по себе одноэлементные наборы не могут дать удовлетворительных результатов, однако на этом примере хорошо видно, как может меняться картина при переходе от одного признака (в общем случае, одного набора признаков) к другому признаку (набору). В строчках 12 и 14 помечены столбцы, соответствующие классификатору 3, где результаты классификации объектов классов 1 и 3 противоположны. Аналогичные результаты наблюдаются и в других строчках. Имеются также и довольно «приличные» (для одноэлементных наборов) результаты распознавания на уровне 0.6 и даже 0.7.

Далее рассмотрим наборы, состоящие из 15 признаков (см. рис. 2, справа). Каждый такой набор представляет собой полный набор признаков, из которого исключен один признак. Из рис. 2 видно, что наилучшие результаты дает классификатор 3, а классификатор 2 дает наихудшие результаты.

Аналогичную картину можно получить и для сочетаний, состоящих их двух признаков (из всех признаков, исключая какие-либо два признака), и для сочетаний, состоящих из трех признаков (из всех признаков, исключая какие-либо три признака), и т. д. «Большие» наборы, очевидно, дают лучшие результаты при распознавании. Однако имеющиеся между признаками статистические зависимости или наличие в матрицах данных «шума» (признаки, не несущие информации об имеющихся классах) делают «большие» наборы ненадежными. Поэтому встречаются и «малые» наборы, дающие хорошие результаты при распознавании.

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

9.2. Алгоритм `Add`

На начальном этапе в алгоритме `Add` последовательно рассматриваются *одноэлементные* наборы и выбирается такой одноэлементный набор, который дает наилучшие результаты классификации. Затем последовательно рассматриваются уже двухэлементные наборы, которые содержат полученный на предыдущем шаге оптимальный одноэлементный набор, и выбирается оптимальный *двухэлементный* набор. Далее повторяется тоже самое, но уже для *трехэлементных* наборов.

Алгоритм прекращает свою работу, когда достигается требуемое количество признаков. Если встречаются равноценные наборы, то выбирается *любой* из них и работа алгоритма продолжается. Можно параллельно рассматривать все встречающиеся альтернативы, однако это существенно увеличит объем вычислений, нивелируя все достоинства метода.

9.3. Алгоритм `Add*`

Модифицированный алгоритм `Add` последовательно добавляет признаки до тех пор, пока удается улучшить качество классификации. Если улучшить качество классификации не удается, то алгоритм аварийно завершает свою работу. Основное преимущество модифицированного алгоритма в том, что отсеивается большое количество заведомо неоптимальных наборов, а получаемые в результате наборы обладают экстремальными свойствами.

9.4. Алгоритм `Del`

Алгоритм последовательного исключения `Del` действует по аналогичной схеме: последовательно рассматриваются наборы, в которых *отсутствует* ровно один признак из числа тех, которые присутствуют в исходном (начальном) наборе, и среди полученных наборов выбирается наилучший. Начиная с «больших» наборов, где число ошибок классификации в среднем невелико, алгоритм переходит к «малым» наборам, где число ошибок становится больше. Однако при наличии между признаками статистических зависимостей уменьшение числа признаков в наборе приводит к исключению таких зависимостей.

9.5. Алгоритм `Del*`

Модифицированный алгоритм `Del` действует аналогично алгоритму `Add*` с той лишь разницей, что признаки исключаются, а не добавляются.

9.6. Алгоритм `AddDel`

Смешанная стратегия последовательного перебора состоит в попеременном применении алгоритмов `Add` и `Del`. Для этого задается порядок выполнения и количество шагов каждого алгоритма. В этом случае появляется шанс исключить ошибочно добавленные признаки или добавить ошибочно исключен-

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

ные признаки, что позволяет приблизиться к оптимальному набору, однако может существенно увеличиться время поиска.

9.7. Алгоритм `Prob`

Алгоритм `Prob` состоит в том, что признакам назначаются определенные веса, которые и влияют на отбор признаков. В начале веса одинаковы. Затем случайным образом выбирается требуемое количество признаков m и проводится классификация. Эта операция повторяется несколько, скажем, r раз. Выбирается наилучший и наихудший из полученных r наборов. Далее производится пересчет признаков, причем так, чтобы признак, вошедший в наилучший набор, получает больший вес, а признак, вошедший в наихудший набор, получает меньший. (Признаки, попавшие в оба набора, сохраняют свои прежние веса.) Повторяя указанную процедуру несколько раз, можно придти к тому, что будут выбираться одни и те же признаки, что свидетельствует об окончании работы алгоритма. Получаемые наборы оказываются близкими к оптимальным.

При реализации алгоритма используется датчик псевдослучайных чисел, равномерно распределенных на единичном интервале, который разбивается на m отрезков с длинами, соответствующими весам признаков (при некоторой заранее заданной нумерации признаков). Изменению весов соответствует изменение длин интервалов. Возможно уменьшение или увеличение длины интервала на некоторую положительную величину h. Этот параметр управляет скоростью «сходимости» к оптимальному набору. Если этот параметр существенно больше нуля (близок к 0.5), то число шагов алгоритма будет небольшим, а результирующий набор неоптимальным (возможны исключения). Если этот параметр близок к нулю, то можно довольно долго приближаться к оптимуму. Поэтому, выбор этого параметра — результат компромисса, основанного на подробном исследовании.

Число r просматриваемых вариантов на каждом шаге также является управляющим параметром, выбор которого требует глубокого исследования.

Основной недостаток метода СПА заключается в том, что этот метод позволяет (в лучшем случае) получать близкие к оптимальным наборы для заданного числа признаков, но ничего не сообщает об оптимальном числе признаков. Для разрешения данной проблемы приходится решать ровно M задач (точнее, меньшее их число, за счёт тех случаев, которые получаются при полном переборе).

10. Заключение

Рассмотренные нами методы отбора информативных признаков обладают как достоинствами, так и недостатками. Каждый из них является «хорошим» в каком-то одном отношении. Метод последовательного перебора не гарантирует, что на каком-либо шаге не будет исключен признак, входящий в оптимальный набор, или, наоборот, будет добавлен признак, не входящий в оптимальный набор. Другая особенность метода — изменение числа признаков, что означает перебор признаковых пространств различных размерностей. Смешанные алгоритмы, однако, предоставляют возможность многократного добавления и/или исключения признаков, что нивелирует недостатки последовательного перебора. А манипуляции с количеством используемых базовых алгоритмов

Труды СПИИРАН. 2007. Вып. 4. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 4. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

и глубиной поиска позволяют конструировать стратегии, учитывающие информацию о текущем состоянии (используемые признаки, статистические зависимости и имеющиеся предпочтения). Метод СПА также предоставляет широкие возможности при варьировании как числа просматриваемых наборов «за раз», так и величины шага, на который меняются веса отдельных признаков. Дополнительное достоинство метода СПА — это то, что мы можем оценить его эффективность, то есть — вероятность встретить среди случайно отобранных наборов признаков оптимальный. Полный перебор вариантов, когда он оказывается осуществимым, предоставляет возможность сравнить результаты, полученные другими методами, и получить необходимые числовые характеристики.

Автор благодарит проф. С. П. Романова (Институт физиологии им. И. П. Павлова, Институт мозга РАН) за предоставленный экспериментальный материал.

Литература

1. *Пчелин М. Г., Романов С. П., Якимовский А. Ф.* Метод тензометрии для количественной оценки тремора // Физиологичесий журнал им. И. М. Сеченова. 1996. Т. 82, № 2. С. 118–123.
2. *Вассерман Е. Л., Карташев Н. К., Полонников Р. И.* Фрактальная динамика электрической активности мозга. СПб.: Наука, 2004. 208 с.
3. *Жвалевский О. В.* О возможности автоматизации диагностики болезни Паркинсона методом анализа фрактальной динамики // Труды СПИИРАН. 2006. Вып. 3, т. 2. СПб.: Наука, 2006. С. 187–297.
4. *Айвазян С. А., Бежаева З. И., Староверов О. В.* Классификация многомерных наблюдений. М.: Статистика, 1974. 240 с.
5. *Воронцов Н. К.* Отбор информативных признаков // Сайт Н. К. Воронцова, раздел «Преподавание»: <http://www.ccas.ru/voron/teaching.html>.
6. *Лбов Г. С.* Выбор эффективной системы признаков // Вычислительные системы. 1965. Вып. 19. Новосибирск: Наука, Сибирское отделение, 1965. С. 21–34.
7. *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. Новосибирск: Изд-во Ин-та математики, 1999. 270 с.