АНАЛИЗ СТРУКТУРНЫХ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
2005 год Выпуск 174

УДК 519.95

АЛГОРИТМ GRAD ДЛЯ ВЫБОРА ПРИЗНАКОВ[1]

Н.Г. Загоруйко, О.А. Кутненко

Уменьшение размерности признакового пространства целесообразно по двум причинам. Во-первых, исключение дублирующих признаков приводит к сокращению объема вычислений в процессе распознавания или поиска ближайшего аналога. И, во-вторых, устранение неинформативных, «шумящих» признаков повышает надежность распознавания.

Для решения задачи выбора подмножества наиболее информативных элементов из их большого исходного множества может быть использован алгоритм AdDel. Он состоит из последовательно сменяющих друг друга процедур добавления (Addition) наиболее информативных и исключения (Deletion) наименее информативных элементов. В разработанном алгоритме GRAD («гранулированный AdDel») метод AdDel работает на предварительно сформированном множестве признаков — гранул (или наборов), состоящих из нескольких исходных элементов. Как и AdDel, предложенный алгоритм позволяет указать состав и наилучшее количество характеристик.

[1]Работа выполнена при финансовой поддержке РФФИ (проект № 05-01-00241).

1. Метод последовательного сокращения признаков (алгоритм Del)

Пусть задача состоит в том, что из N признаков нужно выбрать наиболее информативную систему, состоящую из n признаков. В 1963 г. Т.Мэрилл и О.Грин [1] предложили алгоритм Deletion последовательного сокращения исходного множества признаков за счет вычеркивания на каждом шаге наименее полезных. Алгоритм работает следующим образом. Исключается из системы первый признак и находится ошибка α_{11}, которую дают оставшиеся N–1 признаков. Меняются ролями первый и второй признаки и находится ошибка α_{12} в новом $(N-1)$-мерном признаковом пространстве. Эта операция поочередного исключения одного признака проводится N раз. Среди полученных величин $\alpha_{11}, \ldots, \alpha_{1n}, \ldots, \alpha_{1N}$ находится самая большая. Она укажет на наименее информативный признак. Исключим его и приступим к аналогичному испытанию оставшихся в системе $(N-1)$ признаков, что позволит найти самый неинформативный из них и снизить размерность пространства до $(N-2)$. Эти процедуры повторяются $(N-n)$ раз. Подсистема из n оставшихся признаков будет обладать наибольшей информативностью среди $(n+1)$ проверенных подсистем такой же размерности.

Количество проверяемых систем признаков при этом методе значительно меньше, чем объем полного перебора. Так, даже в простом случае при $N = 50$ и $n = 25$ количество операций алгоритма Del на 12 порядков меньше объема полного перебора.

2. Метод последовательного добавления признаков (алгоритм Ad)

В том же 1963 году Ю.Барабаш и др. [2] предложили алгоритм Addition, который отличается от предыдущего тем, что порядок проверки подсистем признаков начинается не с N-мерного пространства, а с одномерных пространств. Для этого делается распознавание контрольной последовательности по каждому из N признаков в отдельности и в информативную подсистему включается признак, давший наименьшее число ошибок. Затем

к нему по очереди добавляются все (N–1) признаков по одному. Получающиеся двумерные подпространства оцениваются по количеству ошибок распознавания.

В итоге выбирается наиболее информативная пара признаков. К ней таким же путем подбирается наилучший третий признак из оставшихся (N–2) признаков и так продолжается до получения системы из n признаков.

Оба описанных алгоритма дают оптимальное решение на каждом шаге, но это не обеспечивает глобального оптимального решения.

3. Алгоритм AdDel

В 70-е годы многие авторы исследовали различные комбинации этих двух алгоритмов (см. обзор в [3]). Сравнение различных вариантов последовательного включения алгоритмов Addition (Ad) и Deletion (Del) показало преимущества алгоритма AdDel перед алгоритмами Ad, Del и DelAd [4]. Алгоритм AdDel состоит в следующем. Методом Ad набирается некоторое количество (n_1) информативных признаков, затем n_2 из них ($n_2 < n_1$) исключается методом Del. После этого алгоритмом Ad размерность набора информативных признаков наращивается на величину n_1 и становится равной ($2n_1 - n_2$). В этот момент снова включается алгоритм Del, который исключает из системы n_2 «наименее ценных» признака. Такое чередование алгоритмов Ad и Del продолжается до достижения заданного количества признаков n.

Наблюдения показывают, что по мере увеличения числа признаков качество распознавания вначале растет, потом рост прекращается и начинается его снижение за счет добавления малоинформативных, шумящих признаков. Перегиб кривой качества позволяет **указать оптимальное количество признаков**. Это очень важное свойство алгоритмов семейства AdDel, которым не обладают другие алгоритмы выбора информативных признаков.

4. Алгоритм GRAD

Более общий вид алгоритма этого класса описан в [5]. Его отличие от других состоит в том, что добавление n_1 и исключение n_2 признаков делается не по одному, а всеми возможными сочетаниями из имеющихся по n_1 и n_2 соответственно. Полезность такого подхода понятна. Действительно, несколько признаков, каждый из которых не информативен, вследствие особого вида взаимной зависимости могут образовать информативную «гранулу». Мы убедились, в этом на таком простом примере. На генетических данных с исходным числом признаков $N = 319$ выбирались информативные подпространства при постепенном наращивании их размерности. При $n = 2$ были использованы два метода — алгоритм AdDel и метод полного перебора. Оказалось, что полный перебор выявил 6 пар, информативность которых была выше информативности лучшей пары, найденной методом AdDel. В состав все этих более успешных пар входили признаки, индивидуальная информативность которых была ниже информативности самого информативного признака, выбранного алгоритмом AdDel на первом шаге.

Однако переход от поштучного рассмотрения признаков к рассмотрению всех возможных их сочетаний сопряжен с быстрым ростом трудоемкости алгоритма.

В тех же экспериментах было обнаружено, что если все признаки упорядочить по убыванию их индивидуальной информативности, то в составе наиболее информативных пар преобладают признаки с малыми порядковыми номерами. Гипотеза о том, что признаки, обладающие малой индивидуальной информативностью, редко попадают в состав наиболее информативных пар, была проверена и подтверждена в экспериментах на генетической таблице с количеством признаков $N = 5527$ (см. рис. 1).

Из приведенных фактов был сделан следующий вывод: в алгоритмах выбора информативных подсистем нужно применять элементы полного перебора настолько широко, насколько это позволяют машинные ресурсы. В разработанном нами алгоритме GRAD («гранулированный AdDel») метод AdDel работает на множестве G предварительно отобранных наиболее информатив-

ных «гранул», каждая из которых состоит из w признаков, $w = 1, 2, 3 \ldots W$.

Выбор W нами делался исходя из двух соображений. Первое из них основано на учете ограничений на возможности решения реальных переборных задач на вычислительных машинах распространенного типа. Второе основано на гипотезе о преобладании простых закономерностей над сложными. Вполне ве-

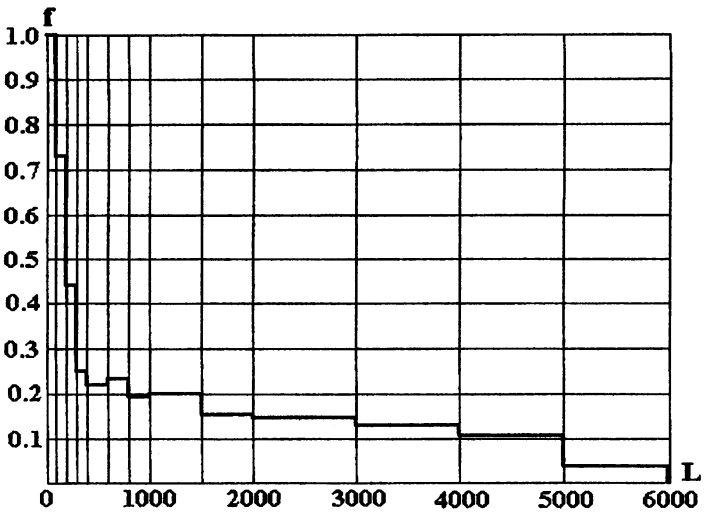

Рис. 1. Зависимость частоты f попадания в информативную подсистему от порядкового номера L индивидуальной информативности

роятно, что среди индивидуально неинформативных признаков может обнаружиться два признака, так дополняющих друг друга, что вместе они образуют высокоинформативную пару. Гораздо труднее представить такую подсистему из большого числа признаков w, чтобы не только каждый из них, но и все парные, тройные и прочие их сочетания, вплоть до подсистем из $(w-1)$

признаков, были бы неинформативными, и лишь добавление одного w–го признака делало бы систему высокоинформативной. В итоге было выбрано значение $W = 3$.

Первый этап работы алгоритма GRAD состоит в следующем. Вначале все N признаков упорядочиваются по индивидуальной информативности в список P_0. При большом значении N полный перебор сочетаний по 2 и по 3 затруднителен. По этой причине формируется список P_1, состоящий из $m_1 < N$ наиболее информативных признаков. Из списка P_1 методом полного перебора формируется список P_2, состоящий из m_2 двухместных гранул, и список P_3, состоящий из m_3 трехместных гранул. Будем считать гранулы мощности 1,2 и 3 вторичными признаками, как это делается в методе группового учета аргументов. Величины m_1, m_2, m_3 задавались, исходя из возможности настольного персонального компьютера, и общее их количество в списке P_4 не превышало нескольких сотен.

5. Сравнение эффективности алгоритмов AdDel и GRAD

Исходная система содержала 5527 генетических признаков, которыми описывались 35 пациентов, разделенных на две группы — образы «здоров» и «болен диабетом» [6]. Признаковые подсистемы мощности $n > 3$ формировались из 300 признаков списка P_4 двумя сравниваемыми методами: AdDel и GRAD. При этом каждый из них работал в двух режимах: «без повторов» и «с повторами». В первом режиме гранула, включенная в создаваемую подсистему признаков, из дальнейшего рассмотрения исключалась. Во втором режиме каждая гранула могла присоединяться к создаваемой подсистеме любое количество раз. Если какая–то гранула включалась в подсистему несколько раз, это означало, что ее вес в подсистеме был больше весов других гранул.

Выбирались подсистемы различной размерности $n < 61$. Их информативность оценивалась по числу правильных результатов при скользящем экзамене обучающей выборки. Количества подсистем W с максимальной информативностью $I = 35$, найденных четырьмя разными версиями алгоритмов, и количество процедур

T, затраченных на поиск одной подсистемы, приведены в таблице. Под процедурой T здесь понимается последовательность операций, связанных с формированием и оценкой информативности одного варианта подпространства.

Таблица

	GRAD	GRAD с повт.	AdDel	AdDel с повт.
W	46880	9862	8683	5101
T	11	55	12	28

Из результатов этих экспериментов можно сделать следующие выводы. Алгоритм GRAD имеет явные преимущества перед другими алгоритмами семейства AdDel по количеству найденных информативных подсистем. Следует иметь в виду, что таблицы реальных данных обычно содержат высокий уровень помех, и опора не на одно, а на несколько наиболее информативных подсистем обладает более высокой помехоустойчивостью. Кроме того, алгоритм GRAD, даже с учетом затрат времени на первый этап формирования гранул, выигрывает также и по времени, затрачиваемому на поиск одной подсистемы.

6. Соотношение числа шагов добавления и вычеркивания

Вопрос о количестве шагов n_1 добавления лучших и исключения n_2 худших признаков (размеры «зубцов») в алгоритмах семейства AdDel остается открытым. Качество рассматриваемых подсистем зависит от неизвестных свойств обучающей выборки (наличие тесно связанных групп признаков, их частота встречаемости и пр.). По этой причине обоснованных аналитических рекомендаций по оптимальному соотношению величин n_1 и n_2 найти не удается. Некоторые практические рекомендации можно получить из опыта применения разных сочетаний n_1 и n_2 при решении ряда задач.

Выяснилось, что выигрывают варианты, в которых соотношение $n_1 : n_2 = 2 : 1$, а среди них лучшие результаты получены при значениях $n_1 = 6$, $n_2 = 3$. Эти параметры были использованы во всех дальнейших экспериментах.

Алгоритм GRAD в окончательном виде состоит из двух этапов.

На **первом этапе** формируется вторичное пространство гранулированных признаков. Для этого:

1) упорядочивается по информативности список P_0 единичных признаков;

2) первые m_1 признаков включаются в список P_1 гранул единичной мощности;

3) методом полного перебора из признаков списка P_1 формируются все парные сочетания. Из них формируется список P_2 состоящий из m_2 наиболее информативных гранул мощности 2;

4) методом полного перебора из признаков списка P_1 формируются все сочетания признаков по 3. Из них формируется список P_3, состоящий из m_3 наиболее информативных гранул мощности 3;

5) в список вторичных признаков P_4 включаются все гранулы из списков P_1, P_2 и P_3.

Второй этап состоит в формировании информативных подпространств из вторичных признаков с помощью алгоритма Ad-Del по схеме «6:3». При этом все вторичные признаки, входящие в список P_4, участвуют в выборе на равных основаниях.

В целесообразности использования нескольких информативных подсистем для **построения коллективных решающих правил** мы убедились на примере тех же экспериментальных данных при решении задачи упорядочения пациентов по шкале «здоров–болен». С этой целью были взяты подсистемы с информативностью от 30 до 34. Каждая из них в отдельности не обеспечивает безошибочного распознавания 35 объектов. Для всех подсистем в режиме скользящего экзамена вычислялось значение функции F принадлежности пациентов к двум распознаваемым классам: $F = 1 - \dfrac{2r_1}{r_1 + r_2}$, где r_1 и r_2 — расстояния до ближайших представителей 1–го и 2–го образов соответственно.

В рассмотрение были включены не только 35 пациентов, входящих в контрастные классы «здоров» и «болен диабетом», но и 8 пациентов из промежуточного класса «высокая группа риска».

Значения функции принадлежности усреднялись по разному количеству найденных подсистем признаков. Если при усред-

нении по нескольким подсистемам (5–10) некоторые пациенты второго промежуточного класса ошибочно размещались среди пациентов первого либо третьего классов, то после усреднения по 51 подсистеме ошибок в порядковых позициях уже не было. На рис. 2 вертикальными линиями указаны границы между тремя классами. Видно, что все пациенты второго класса занимают промежуточные порядковые позиции между первым и третьим классами.

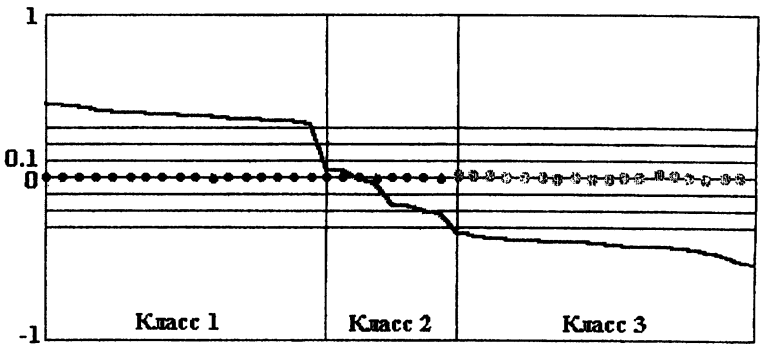

Рис. 2. Расположение пациентов в порядке ухудшения здоровья

Решение задачи упорядочивания пациентов по индексу состояния здоровья позволяет решать важные практические задачи ранней диагностики заболеваний и контроля хода лечения пациентов.

Высокая эффективность алгоритма GRAD подтвердилась при решении и других задач выбора признаков [7].

Литература

1. MERILL T., GREEN O.M. On the effectiveness of receptions in recognition systems //IEEE Trans. Inform. Theory. — 1963. — Vol. IT-9. — P. 11–17.

2. Автоматическое распознавание образов /Ю.Л.Барабаш и др. — Киев. Изд. КВАИУ. — 1963.

3. Методы, критерии и алгоритмы, используемые при преобразовании, выделении и выборе признаков в анализе данных.

— Вильнюс: Институт математики и кибернетики АН ЛитССР. — 1988.

4. ЗАГОРУЙКО Н.Г. Прикладные методы анализа данных и знаний. — Новосибирск: Изд. ИМ СО РАН. — 1999.

5. КУТИН Г.И. Методы ранжировки комплексов признаков. Обзор. //Зарубежная радиоэлектроника. — 1981. — № 9. — С. 54–70.

6. Groop LC. PGC-lalpharesponsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes /Mootha V.K., Lindgren C.M., Friksson K.F. and oth. //Nat Genet. — 2003. — Vol. 34, № 3. — P. 267- 273.

7. Predictive Analysis of Gene Data from Human SAGE Libraries/Alves A., Zagoruiko N., Okun O. and oth.//Proceeding of the Workshop W10 "Discovery Challenge" ECML/PKDD-2005, Porto, Portugal. Edited by Petr Berka and Bruno Cremilleux, Universite de Caen, France, 2005. — P. 60-71.

Поступила в редакцию
8 августа 2005 года