УДК 519.95

ВЫБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ ДИАГНОСТИКИ ЗАБОЛЕВАНИЙ ПО ГЕНЕТИЧЕСКИМ ДАННЫМ

© 2014 г. **Н.Г. Загоруйко, О.А. Кутненко, И.А. Борисова, В.В. Дюбанов, Д.А. Леванов, О.А. Зырянов**

Федеральное государственное бюджетное учреждение науки Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: zag@mail.ru

Поступила в редакцию 28 сентября 2014 г. Принята к публикации 24 октября 2014 г.

В связи с появлением и активным использованием ДНК-микрочипов при решении различных задач в медицине, биоинформатике и молекулярной биологии усилилась потребность в алгоритмах Data Mining, способных обрабатывать задачи, в которых число анализируемых объектов на порядки меньше числа описывающих признаков. Однако большинство из существующих ныне алгоритмов изначально не предназначено для решения подобных сложных, плохо обусловленных задач. Нами разработан подход, основанный на идее конкурентного сходства, который позволяет разрабатывать алгоритмы, лучше приспособленные для этих целей. Одним из таких алгоритмов является предложенный нами алгоритм FRiS-GRAD, который одновременно решает задачу распознавания и задачу выбора системы информативных признаков. Эффективность его работы проиллюстрирована на различных медицинских задачах в сравнении с наиболее популярными алгоритмами выбора информативных признаков и распознавания.

Ключевые слова: экспрессия генов, функция конкурентного сходства, выбор информативных признаков, распознавание.

ВВЕДЕНИЕ

В настоящее время увеличивается количество публикаций с данными об экспрессии генов у пациентов − носителей различных заболеваний. Данные имеют вид таблиц из M объектов (пациентов) и N признаков (экспрессии генов). Один из основных видов анализа таких данных состоит в выборе подмножества признаков, по которым можно было бы делать диагностику заболеваний. Эту задачу можно было бы решить, оценивая информативность каждого из N генов в отдельности и выбирая заданное количество $n < N$ генов с наибольшей индивидуальной информативностью.

Но между генами имеются зависимости, учет которых заставляет выбирать такие гены, которые дополняли бы друг друга и образовывали подмножества генов с максимальной коллективной информативностью. Точное решение этой задачи методом полного перебора всех сочетаний из N по n генов в общем случае получить нельзя. Предлагаются различные эвристические методы выбора информативных характеристик. В работе Jeffery с соавт. (2006) описаны десять наиболее популярных методов выбора признаков. Их типичным недостатком является предположение о том, что признаки независимы. Кроме того, без всякого обоснования выбрано заданное количество n генов.

В данной работе описан алгоритм FRiS-GRAD для выбора информативных признаков, который учитывает взаимные зависимости между признаками и автоматически определяет их оптимальное количество n. Алгоритм основан на использовании тернарной меры сходства между объектами в виде FRiS-функции (Zagoruiko *et al.*, 2008). Это позволяет сделать прозрачным способ построения решающих правил, оценить количественно компактность образов и информативность признаков. Приведены примеры решения реальных генетических задач.

Что такое FRiS-функция?

Главным элементом всех методов анализа данных является мера сходства между объектами или признаками. Считать, что для оценки сходства между объектами *a* и *b* достаточно знать расстояние *r*(*a, b*) между ними, неправильно. Можно ли при *r*(*a, b*) = 5 считать, что объект *a* похож на объект *b* настолько, чтобы их можно было включить в один класс? А Москва от Санкт-Петербурга далеко или близко? Ответить на такие вопросы можно, только зная ответ на вопрос «По сравнению с чем?»

Следовательно, надо знать не только расстояние *r*(*a, b*), но и расстояние *r*(*a, c*) до объекта *c*, который является ближайшим к *a* конкурентом объекту *b*. Сходство объекта *a* с объектом *b* в конкуренции с объектом *c* оценивается по формуле

$$F(a,b \mid c) = \frac{r(a,c) - r(a,b)}{r(a,c) + r(a,b)}. \quad (1)$$

Если объекты *a* и *b* совпадают, то их сходство равно 1. Если расстояния от *a* до *b* и *c* одинаковы, то сходство равно 0. Если же *a* совпадает с *c*, то сходство *a* с *b* равно –1.

Как построены решающие правила?

Используется решающее правило прецедентного типа. Из M_i объектов a_i обучающей выборки каждого *i*-го образа выбирают типичные объекты («столпы»). Столпом назначается такой объект a_i, сходство с которым всех остальных объектов a_j данного образа в конкуренции с ближайшими объектами b_j чужого образа максимально (Zagoruiko *et al.*, 2008). Вокруг каждого выбранного столпа формируется кластер. В него входят объекты, сходство которых *F* со столпом выше порога, например, $F > 0$. Если какие-то объекты не вошли в кластеры (оказались незащищенными), то среди них выбирается следующий столп. Им становится объект любого из *K* образов, сходство с которым остальных незащищенных объектов этого образа в конкуренции с любым ближайшим объектом чужого образа максимально. Такая процедура последовательного увеличения столпов продолжается, пока все *M* объектов не окажутся включенными в кластеры.

В итоге формируется список из *k* столпов с указанием количества объектов, которые входят в их кластеры. Решающее правило для распознавания принадлежности контрольного объекта *z* к одному из *K* образов состоит в следующем. Вычисляются расстояния от *z* до всех *k* столпов. Выбираются два самых близких столпа, принадлежащие разным образам. Объект *z* считается принадлежащим тому образу, на столп которого он похож больше всего. Величина FRiS-функции показывает надежность принятого решения.

Как оценить компактность?

Компактность *i*-го кластера C_i равна сумме сходств всех M_i объектов a_j кластера со своим столпом s_i в конкуренции с ближайшими столпами s_v других образов (Загоруйко и др., 2010):

$$C_i = \sum_{j=1}^{M_i} F(a_j, s_i \mid s_v). \quad (2)$$

Если для описания обучающей выборки *K* образов потребовалось использовать *k* столпов, то компактность *C* описания *M* объектов выборки равна

$$C = \frac{1}{M} \sum_{i=1}^{M} C_i. \quad (3)$$

Алгоритм выбора столпов

1.	Для всех a_i, $a_j \in M_i$
1.1.	вычислить $S_i = \sum F(a_i, a_j \mid b_j)$.
2.	Объект $a_i = \text{argmax}\{S_i/i = 1, ..., M_i\}$ назначается столпом s_i.
3.	Повторить пункты 1 и 2 *K* раз.
4.	Сформировать *K* кластеров.
5.	Если все объекты входят в кластеры, то конец.
6.	Если вне кластеров есть $M' < M$ объектов, для них повторить пункты 1–5.

Уклонение от переобучения

С ростом числа столпов сумма компактностей C_i кластеров монотонно увеличивается. Но мощность M_i очередных кластеров обычно меньше предыдущих. Наступает такой момент, когда появляются кластеры, состоящие из одного или нескольких объектов. Наличие таких кластеров свидетельствует о наступлении стадии переобучения. Для обнаружения этого момента перехода от обучения к переобучению используется функция Q качества описания K образов, которая тем больше, чем больше компактность C и чем меньше количество столпов k:

$$Q = C\,\frac{K}{k} \qquad (4)$$

Наличие штрафа за превышение количества столпов k над количеством образов K приводит к тому, что функция Q сначала растет, затем начинает снижаться. Точка перегиба функции $Q = f(k)$ указывает на момент, когда процесс наращивания числа столпов нужно остановить. Объекты, которые к этому моменту не вошли ни в один кластер, не отражают основные закономерности распределения образов и из дальнейшего использования исключаются (цензурируются). Эксперименты на большом числе модельных задач показали, что цензурируются обычно 12–15 % обучающей выборки. Ошибка распознавания контрольной выборки уменьшается в результате цензурирования в 1,5–2,0 раза (Загоруйко, 2013).

Как выбрать признаки?

Исходные данные часто содержат признаки, которые не несут полезную информацию для решения конкретной задачи. Нужно выбрать такое подмножество признаков (в нашем случае – генов), которые необходимы и достаточны для диагностики заданного заболевания.

Известны «жадные» алгоритмы выбора признаков Addition (Ad), когда на каждом шаге к имеющимся признакам добавляется самый полезный, и Deletion (Del), когда из имеющихся признаков исключается самый бесполезный признак. Оба этих алгоритма локально оптимальны. Чтобы уклониться от попадания в локальный оптимум, используют комбинированную процедуру AdDel, в которой чередуются этапы наращивания подсистемы на $n1$ признаков с процедурой сокращения подсистемы на $n2$ признаков, $n2 < n1$.

В процессе увеличения размерности подсистемы такой процедурой «два шага вперед – один назад» информативность подсистемы растет, затем рост останавливается и начинается уменьшение информативности. Точка перегиба функции информативности указывает на оптимальное количество n признаков.

Можно добавлять и исключать не отдельные признаки, а гранулы, состоящие из двух или трех признаков. Самые информативные пары и тройки признаков можно находить методом полного перебора.

На этом основан алгоритм выбора информативных признаков FRiS-GRAD (гранулированный AdDel), который мы использовали при решении разных задач, в том числе задач с генетическими данными (Загоруйко, 2013).

Решение задачи диагностики лейкемии

Особенность генетических задач заключается в том, что количество признаков (генов) велико: тысячи, десятки тысяч. Это на два – три

Алгоритм выбора признаков

1. Для всех $j = 1, \ldots, N$ признаков вычислить компактность C_j.
2. Признак $x_j = \text{argmax} C_j\{C_j/j = 1, ..., N\}$ внести в подсистему.
3. Повторить пункты 1 и 2 $n1$ раз.
4. Оценить компактность C'' подсистемы.
5. Признак x_j', без которого получается $C''_{\max}$, исключить из подсистемы.
6. Повторить пункты 4 и 5 $n2$ раза, $n2 < n1$.
7. Для признаков, не входящих в подсистему, повторить пункты 2–6.
8. Если на i-м и $(i + 1)$-м шагах $C''_{i+1} < C''_i$, то конец.

порядка больше количества объектов (пациентов). Одна из задач состояла в выборе подмножества генов, по экспрессии которых можно было бы отличать друг от друга пациентов с двумя типами лейкемии – ALL и AML (Guyon *et al.*, 2002). Обучающая выборка содержала 38 объектов, тестовая – 34 объекта. Исходное количество признаков (генов) $N = 7\,129$.

Результаты решения этой задачи, описанные в работе (Guyon *et al.*, 2002), таковы. Информативное подмножество признаков выбиралось методом RFE (разновидностью алгоритма Deletion), решающие правила основаны на методе опорных векторов SVM (Vapnik, 1998). Были найдены наилучшие подсистемы, размерность которых кратна степени числа 2: 4 096, 2 048, …, 4, 2 и 1. По двум лучшим признакам, которые можно выбрать по результатам обучения, правильно распознано 30 объектов из 34, по четырем лучшим признакам – 31, по 128 признакам – 33 объекта (табл. 1).

Нами на тех же данных получены следующие результаты. Информативное подмножество признаков выбрано с помощью алгоритма FRiS-GRAD. Информативность признаков оценена по критерию FRiS-компактности. Из 7 129 признаков выбрано 18 признаков, из которых программа FRiS-Stolp построила 30 вариантов решающих правил. В состав каждого правила входит с разными весами от трех до шести признаков. Первые 10 правил показаны в табл. 2.

Первые 27 правил из 30 дают результат 34 из 34. Различия между приведенными результатами могут зависеть как от метода выбора признаков, так и от типа решающих правил. Для сравнения решающих правил SVM и FRiS был проведен такой эксперимент.

В подпространстве двух признаков (генов 803 и 4846), выбранных методом RFE, по правилу SVM получено 30 правильных ответов, а FRiS-методом – 33.

По лучшему одному гену (4846), выбранному методом RFE, результат SVM равен 27, а результат FRiS равен 30. А по лучшему гену (2461), выбранному алгоритмом GRAD, метод FRiS дает 32 правильных ответа (табл. 3).

Отсюда можно сделать вывод, что как метод выбора признаков, так и решающие правила, основанные на FRiS-функции, обладают высокими конкурентными качествами.

Таблица 1

Результаты обучения и контроля при выборе признаков методом RFE и решающем правиле SVM (обучающая выборка 38 объектов, тестовая выборка 34 объекта)

Число признаков	Критерий выбора	Распознано правильно
7 129	0,85	29
4 096	0,71	24
2 048	0,85	29
1 024	0,94	32
512	0,88	30
256	0,94	32
128	0,97	33
64	0,94	32
32	0,97	33
16	1,00	34
8	1,00	34
4	0,91	31
2	0,88	30
1	0,79	27

Таблица 2

Выбор признаков методом FRiS-GRAD, решающие правила FRiS-Stolp

FRiS	Решающие правила	P
0,72656	537/1,1833/1,2641/2,4049/2	34
0,71373	1454/1,2641/1,4049/1	34
0,71208	2641/1,3264/1,4049/1	34
0,71077	435/1,2641/2,4049/2,6800/1	34
0,70993	2266/1,2641/2,4049/2	34
0,70973	2266/1,2641/2,2724/1,4049/2	34
0,70711	2266/1,2641/2,3264/1,4049/2	34
0,70574	2641/2,3264/1,4049/2,4446/1	34
0,70532	435/1,2641/2,2895/1,4049/2	34
0,70243	2641/2,2724/1,3862/1,4049/2	34

Таблица 3

Результаты распознавания двумя решающими правилами SVM и FRiS-Stolp по двум лучшим признакам, выбранным методом RFE

Метод	Best features	SVM	FRiS-Stolp
RFE	803,4846	30 (88 %)	33 (97 %)
	4846	27 (79 %)	30 (88 %)

Таблица 4

Результаты сравнения FRiS-методов с лучшими результатами, полученными сорока наиболее известными методами

Задача	ALL1	Leuk	Prost	DLBCL	Colon	ALL4	Myel	ALL3	ALL2
Признаки	12 625	7 129	12 625	7 129	2 000	12 625	12 625	12 625	12 625
Объекты $m1/m2$	95/33	47/25	50/53	58/19	22/40	26/67	36/137	65/35	24/91
Рекорды из 40	100,00	95,85	90,19	94,30	88,60	82,06	82,90	59,58	78,23
FRiS	100,00	100,00	96,3	96,9	95,6	88,2	84,8	87,6	85,6
Рейтинг FRiS	1	1	1	1	1	1	1	1	1

Сравнение с наиболее известными методами выбора признаков

В работе Jeffery с соавт. (2006) проведено сравнение десяти наиболее известных методов выбора признаков на основе результатов решения девяти задач диагностики по генетическим данным. Для каждой выбранной системы признаков строились решающие правила четырех наиболее известных типов. Для каждой из девяти задач было получено сорок различных решений.

Мы выбрали лучшие из них (рекорды) и сравнили их с результатами, полученными комбинацией алгоритма выбора признаков FRiS-GRAD с алгоритмом построения решающего правила FRiS-Stolp (табл. 4). В таблице показаны имена задач, размерность признакового пространства N, количества объектов первого ($m1$) и второго ($m2$) образов и две строки результатов. В последней строке показано место, занятое результатами решения всех девяти задач FRiS-методами.

Для каждой задачи по результату, полученному каждым методом, можно указать его рейтинг: лучший результат занимает первое место, худший – десятое. Если просуммировать места, занятые методом на всех задачах, то можно определить его общий рейтинг. Результаты таких подсчетов представлены в табл. 5, в последней строке которой показана сумма рейтинговых мест, занятых FRiS-методом. Такой же анализ был проведен и по четырем использованным решающим правилам. Его результаты показаны в табл. 6, в которой, как и в табл. 5, чем меньше сумма рейтинговых мест, тем лучше.

Таблица 5

Сумма рейтинговых мест, занятых методами выбора признаков

Метод выбора признаков	Рейтинг
Fold change	47
Between group analysis	43
Analysis of variance (ANOVA)	43
Significance analysis of microarrays	42
Rank products	42
Welch t-statistic	39
Template matching	38
Area under the ROC curve	37
MaxT	37
Empirical Bayes t-statistic	32
FRiS-GRAD	9

Таблица 6

Сумма рейтинговых мест, полученных решающими правилами

Решающее правило	Рейтинг
Between group analysis (BGA)	35
K-nearest neighbours (kNN)	32
Naive bayes classification (NBC)	25
Support vector machines (SVM)	19
FRiS-Stolp	9

ЗАКЛЮЧЕНИЕ

По приведенным результатам можно сделать вывод о высокой эффективности сочетания алгоритмов выбора признаков FRiS-GRAD и построения решающих правил FRiS-Stolp для решения сложных задач диагностики по генетическим данным.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке РФФИ по грантам 11-01-00156 и 14-01-00039 и интеграционным проектам СО РАН № 54 и 87.

ЛИТЕРАТУРА

Загоруйко Н.Г. Когнитивный анализ данных. Новосибирск: Академическое издательство ГЕО, 2013. 186 с.

Загоруйко Н.Г., Борисова И.А., Дюбанов В.В., Кутненко О.А. Количественная мера компактности и сходства в конкурентном пространстве // Сибирский журнал индустриальной математики. Новосибирск, 2010. Т. 13. № 1 (41). С. 59–71.

Guyon I., Weston J., Barnhill S., Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines // Machine Learnin. 2002. V. 46 (1–3). P. 389–422.

Jeffery I., Higgins D., Culhane A. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data // BMC Bioinformatics. 2006. V. 7. P. 359.

Vapnik V.N. Statistical Learning Theory. Wiley-Interscience, 1998.

Zagoruiko N.G., Borisova I.A., Dyubanov V.V., Kutnenko O.A. Methods of Recognition Based on the Function of Rival Similarity // Pattern Recognition Image Analisys. 2008. V. 18. No. 1. P. 1–6.

FEATURE SELECTION IN THE TASK OF MEDICAL DIAGNOSTICS ON MICROARRAY DATA

N. G. Zagoruiko, O. A. Kutnenko, I. A. Borisova, V. V. Dyubanov, D.A. Levanov, O.A. Zyranov

Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia,
e-mail: zag@math.nsc.ru

Summary

In tasks of modern biology, the numbers of attributes often exceed the numbers of objects by orders of magnitude. For the solution of such tasks, a Data Mining method based on using a new measure of similarity between objects in the form of the Function of Rival Similarity (FRiS) is offered. On this basis, methods of quantitative estimation of compactness of patterns, construction of decision rules, and feature selection are developed. All these techniques are implemented in the FRiS-GRAD algorithm. The high efficiency of the algorithm is illustrated by results of solving the task of disease recognition on a microarray dataset.

Key words: gene expression, function of rival similarity, feature selection, pattern recognition.