УДК 57.087.1

К.А. Кузнецов, И.В. Мозговая, К.В. Шегеда

УМЕНЬШЕНИЕ ПРОСТРАНСТВА ПРИЗНАКОВ В ЗАДАЧЕ БИНАРНОЙ КЛАССИФИКАЦИИ

Аннотация. Рассматривается проблема выбора информативных генов в задачах медицинской диагностики. Предлагается новый подход к решению поставленной задачи, включающий в себя фильтрацию, локальный поиск в пространстве признаков, и машины опорных векторов в качестве классификаторов. Приводится результат применения предложенного алгоритма к реальным массивам данных.

Ключевые слова. Выбор признаков, бинарная классификация, обработка многомерных данных, интеллектуальный анализ данных

Постановка задачи. В работе рассматривается следующая задача бинарной классификации. Пусть заданы две группы объектов - g_1 и g_2. Для наблюдения доступна обучающая выборка из $n = n_1+n_2$ объектов (n_1 из g_1, и n_2 из g_2). Каждый объект описывается p-мерным вектором вещественных признаков $x = (x_1, \ldots, x_p)$. В рассматриваемых нами приложениях $p>>n$, т. е. существуют незначимые или избыточные признаки, которые усложняют процесс обучения и ведут к неправильной классификации. Целью является построение модели классификации $f(x)$, на наиболее информативных признаках.

Таким образом, исходными данными задачи являются:

1) $G=\{g_1,g_2,\ldots g_p\}$ – множество признаков (генов, степень выраженности которых можно измерять у пациента). При современных технологиях измерения величина $p\geq 50000$.

2) Вещественная матрица X, элементы X_{ij} которой можно интерпретировать как степень выраженности j-го гена у i-го пациента ($i=1,\ldots,n$, $j=1,\ldots,p$).

3) Бинарный вектор Y, где $Y_i=\{-1,1\}$ ($i=1,\ldots,n$), который представляет собой результат применения некоторого варианта лечения к каждому пациенту.

Считается, что результат лечения целиком определяется генетической предрасположенностью пациентов. Однако, неизвестно ка-

 ISSN 1562-9945

кие именно из генов являются ответственными за результат. Целью задачи является уменьшение количества информативных генов $G_{best} \subseteq G$ и построение модели $Y'=f(X, G_{best})$, с минимальным значением некоторой функции потерь $L(Y,Y')$.

Обзор публикаций. Использование фильтров является наиболее часто используемой методологией в датамайнинге. В [1,2] для выбора признаков использовалась t-статистика и критерий Фишера, а информативными считались гены с наибольшим значением соответсвующей статистики. Недостатком этого подхода является тот факт, что отбор осуществляется только на основе знаний об обучающей выборке на стадии препроцессинга и никакая информация о результатах работы алгоритма в учет не берется.

Альтернативой является идея о включении процесса выбора признаков в алгоритм обучения с учителем. Выбор признаков выполняется так, что любые комбинации признаков оцениваются по их эффективности в конкретном алгоритме классификации. В [1] рассматриваются метод Монте Карло и процедура последовательного прямого выбора (stepwise forward selection). В [2] представлен алгоритм, в котором признаки исключаются рекурсивно в процессе обучения последовательности машин опорных векторов. Данные подходы являются трудоемкими из-за включенного перебора признаков. Кроме того, их стабильность обычно ниже, чем у обычных фильтров.

В [3] было предложено использовать логистическую регрессию для выбора признаков и построения классификатора. Для построения классификатора в [4] были рассмотрены окна Парцена, линейный дискриминант Фишера и деревья решений. Т.к. размерность задачи очень высока, был сделан вывод о том, что использование машины опорных векторов является одним из лучших способов решения проблемы.

Целью настоящей работы является построение алгоритма выбора значимых признаков, сочетающего фильтрацию с интеллектуальным перебором признаков на базе машины опорных векторов.

Критерии оценки качества. Для оценки качества классификации модели используется BCR (balanced classification rate):

$$\mathrm{BCR} = \frac{1}{2} \cdot \left(\frac{\mathrm{T_{NEG}}}{\mathrm{NEG}} + \frac{\mathrm{T_{POS}}}{\mathrm{POS}} \right) \qquad (1)$$

где T_{POS} – число правильно предсказанных позитивных исходов, T_{NEG} – число правильно предсказанных негативных исходов, POS – общее число позитивных исходов, NEG – общее число негативных исходов.

Для оценки стабильности модели после k экспериментов вычисляется индекс Кунчевой по следующей формуле:

$$KI(S_1,...,S_k) = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{\left|S_i \cap S_j\right|}{\left|S_i \cup S_j\right|} \qquad (2)$$

где S_i – множество полученных в ходе i-го эксперимента признаков.

Алгоритм решения задачи

Выбираем количество экспериментов k.

For i=1,k

Случайным образом разделяем выборку X на обучающую (X_L^i) и тестовую (X_T^i) части. При этом $\left|X_L^i\right| = 0.9 \cdot \left|X\right|$.

Проводим выбор наиболее информативных признаков G_{best}^i на части X_L^i и осуществляем построение модели $f(X_L^i, G_{best}^i)$.

Оцениваем качество модели, вычисляя значение $Q_i \leftarrow BCR(f(X_T^i, G_{best}^i))$.

End

Вычисляем $\overline{Q} \leftarrow \frac{1}{k} \cdot \sum_{i=1}^{k} Q_i$ и $KI(G_{best}^1, ..., G_{best}^k)$.

Опишем более подробно процесс выбора наиболее информативных признаков и построения модели.

Алгоритм выбора признаков

For j=1,N

Случайным образом выбираем Xf⊆XLi . При этом $\left|X_f\right| = 0.9 \cdot \left|X_L^i\right|$. Выполняем ранжирование признаков G на выборке Xf. Пусть множество Gj состоит из S лучших генов множества G.

End

Находим $G^* = \bigcap_{j=1}^{N} G_j$, и G+ – признаки, встретившиеся не менее, чем в 10% списков Gj. Положим Gbest←G* и строим модель f(XLi, Gbest). Вычисляем maxBCR←BCR(f(XLi, Gbest)). Пусть exit←false.

While exit=false

exit←true

For each t∈G+

$G_{temp} = G_{best} \cup \{t\}$

Строим модель f(XLi, Gtemp) и вычисляем curBCR←BCR(f(XLi,Gtemp))

If curBCR>maxBCR then

Gbest←Gtemp, maxBCR←curBCR, G+←G+\{t}, exit←false

 ISSN 1562-9945

End
End
return Gbest

Мощность множеств G^* и G^+ зависит от N и применяемого критерия ранжирования

Критерии ранжирования признаков:

1) t-статистики Стьюдента:

$$t_g = \frac{m_+ - m_-}{\sqrt{\frac{S_+}{n_+} + \frac{S_-}{n_-}}} \quad (3)$$

где t_g – оценка гена, $m_+(m_-)$ – среднее значение гена по позитивным (негативным) исходам, $S_+(S_-)$ – дисперсия значения гена по позитивным (негативным) исходам, $n_+(n_-)$ – количество позитивных (негативных) исходов.

2) "Сжимающая" t-статистика ("Shrinkage" t-statistic) предложенная в [6]:

$$t_g = \frac{m_+ - m_-}{\sqrt{\frac{\upsilon_+}{n_+} + \frac{\upsilon_-}{n_-}}}, \quad (4)$$

$$\upsilon_+ = \lambda * \overline{S}_+^2 + (1-\lambda)S_+ \quad (5)$$

где S_+ - дисперсия позитивных значения гена для позитивных исходов, $\overline{S}_+^2$ - среднее значение S_+ по всем генам.

$$\lambda = \mathbf{min}\left(1, \frac{\sum_{k=1}^{p} Var(S_+^k)}{\sum_{k=1}^{p} (S_+^k - \overline{S}_+^2)^2}\right) \quad (6)$$

где p – количество признаков, S_+^k – дисперсия k-го признака для позитивных исходов.

$$Var(S_+^k) = \frac{n_+}{(n_+ - 1)^3} \sum_{i=1}^{n_+} (w_{ik} - \overline{w_k})^2$$
$$w_{ik} = (x_{ik} - m_+^k)^2 \quad (7)$$
$$\overline{w_k} = \frac{1}{n_+} \sum_{i=1}^{n_+} w_{ik}$$

x_{ik} – значение k-го признака i-го пациента, m_+^k – среднее значение k-го признака по позитивным исходам. При вычислении υ_- вместо позитивных исходов рассматриваются негативные.

Построение классификатора. Для построения модели мы предлагаем использовать машину опорных векторов [6] со следующими функциями ядра:

Линейная: $K(x_i, x_j) = x_i^T x_j$

Полиномиальная: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$

Радиальные базисные функции: $K(x_i, x_j) = e^{(-\gamma \|x_i - x_j\|^2)}$

Результаты. В статье рассмотрены два набора данных:

1) *Alon* (см. [7]), содержащий степени выраженности 2000 генов у 62 пациентов. У 40 из них диагностирован рак прямой кишки.

При использовании нашего подхода для данной выборки с параметрами S=64 и N=1000 оказалось, что 18 признаков входят в список G^* всегда, 1790 признаков в этот список не попадают ни разу (соответственно 210 – хотя бы один раз). Таким образом удалось существенно сократить размерность задачи с 62*2000 до 62*210. Если оставить только те признаки, которые входят в «топ-64» не менее, чем в 10% всех ресемплингов, то размерность можно еще уменьшить с 210 признаков до 85. Результаты работы алгоритма с использованием различных критериев и функций ядра представлены в Таблице 1

Таблица 1

Результаты для Alon

	BCR	KI	Отобрано признаков
t-статистика с линейной функцией ядро	0,6733	0,6822	32-37
Shrinkage t-statistic с линейной функцией ядро	0,775	0,6722	31-36
Shrinkage t-statistic с квадратической функцией ядро	0,825	0,64	29-38
Shrinkage t-statistic с радиальными базисными функциями в качестве функции ядро	0,9	0,7444	30-33

2) *LymphomaNoNa* (см. [8])– этот набор данных содержит степени выраженности генов, измеренные по технологии cDNA microarray (Lymphochip). В наборе есть информация о 4026 генов для 47 пациентов из них 24 пациента с B-подобной диффузной B-крупноклеточной лимфомой центров размножения и 23 с активированной B-подобной диффузной B-крупноклеточной лимфомой.

Результаты работы алгоритма для набора данных LymphomaNoNa представлены в Таблице 2

 ISSN 1562-9945

Таблица 2

Результаты для LymphomaNoNa

	BCR	KI	Отобрано признаков
Shrinkage t-статистика с радиальными базисными функциями	0,9333	0,7106	30-37
t-statistic с радиальными базисными функциями	0,9583	0,6471	28-39
Shrinkage t-statistic с радиальными базисными функциями и пороговым значением 40%	0,95	0,7049	28-42

Выводы. В данной работе рассмотрена задача уменьшения пространства признаков в бинарной классификации. Проанализированы существующие методы решения и предложен новый подход для выбора информативных признаков, сочетающий в себе преимущества фильтра и локального поиска. Были проанализированы результаты работы машины опорных векторов с разными функциями ядра в качестве классификатора. Имеются перспективы улучшения подхода с помощью использования более интеллектуальных алгоритмов поиска в пространстве признаков.

ЛИТЕРАТУРА

1. Xiong M, Wuju L., Zhao J. et. al. Feature (gene) selection in gene expression-based tumor classification.//Mol. Gent. Metab.-2001.-Vol.73.-p. 239-247.

2. Furey T., Cristianini N., Duffy N. et. al. Support vector machine classification and validation of cancer tissue samples using microarrya expression data. // Bioinformatics.-2000.-Vol. 16.-p. 906-914.

3. Shevade S., Keerthi S. A simple and efficient algorithm for gene selection using sparse logistic regression.//Bioninformatics.-2003.-Vol.19.-p.2246-2253.

4. Brown M., Grundy W., Lin D. et. al. Knowledge based analysis of microarray gene expression data using support vector machine. // Proc. Natl. Acad. Sci. USA.-2000.-Vol. 97.-p. 262-267.

5. Opgen-Rhein R., Strimmer K. Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach. // Statistical Applications in Genetics and Molecular Biology.-2007.-Vol. 6, Issue 1.-p. 7-8.

6. Хайкин С. Нейронные сети. Полный курс.– М.: ООО «И. Д. Вильямс», 2006. – 1104 с.

7. Alon U. 'Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues'. //[Электронный ресурс]. URL: http://genomics-pubs.princeton.edu/oncology/affydata/index.html

8. Alizadeh A.A. 'Distinct types of Diffuse Large B-Cell Lymphoma Identified By Gene Expression Profile'. [Электронный ресурс]. URL: http://llmpp.nih.gov/lymphoma/data.shtml