ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ
1973 год Выпуск 55

УДК 681.3.06:621.391

ОБ ОЦЕНКЕ КАЧЕСТВА РЕШАЮЩЕГО ПРАВИЛА НА ОСНОВЕ МАЛОЙ ОБУЧАЮЩЕЙ ВЫБОРКИ

Г.С. Лбов, А.Н. Манохин

§ 1. Постановка задачи

Определение прогнозирующей способности решающего правила по выборке, на основе которой строилось само решающее правило, является важной задачей в распознавании образов. При построении решающего правила необходимо учитывать эффект, вызываемый ограниченностью экспериментального материала. Так, например, может оказаться, что квадратичное решающее правило хуже, чем линейное, либо линейное правило, заданное на всем исходном множестве признаков, может оказаться хуже, чем линейное правило, заданное на подмножестве этих признаков.

Этой важной теме посвящены работы [1,2,3,4,6], в которых указанная задача решается в рамках статистической модели. Предполагается, что реализации обучающей выборки независимы и извлекаются из генеральной совокупности случайным образом в соответствии с некоторым распределением $\rho(x,\omega)$, где x -вектор наблюдений признаков $X_1,\dots,X_z$, ω - номер образа.

В работе [1] рассматривается случай нормальных распределений с единичными матрицами ковариации, но с различными векторами математических ожиданий для разных образов. Обобщение на случай равных и неравных матриц ковариаций приводится в работе [2]. Вапником и Червоненкисом [3,4] получен ценный теоретический результат: чем сложнее решающее правило (чем выше так называемая его емкостная характеристика), тем больше требуется объем выборки для того, чтобы получаемое по выборке решающее правило было близко по вероятности ошибочной классификации к оптимальному правилу. Однако эти результаты верны для любой функции распределения $\rho(x,\omega)$, а значит, они должны быть верны и для самого худшего распределения, для которого требуется большой объем обучающей выборки. Оценка необходимого объема выборки получается слишком завышенной.

Решаемая в данной работе задача определения оценки качества решающего правила характеризуется следующими особенностями:

1. Распределение $\rho(x,\omega)$ считается неизвестным, и каких-либо предложений о виде этого распределения не делается.

2. Рассматривается случай, когда объекты различных образов описываются признаками, замеренными в шкале наименований. Вообще говоря, результаты данной работы будут верны и для более общего случая (для признаков, замеренных в различных шкалах). Для этого необходимо перевести разнотипные признаки в признаки наименований. Один из алгоритмов такого перевода приводится в работе [5].

3. Объем выборки N незначителен по сравнению с числом признаков z и тем более по сравнению с числом возможных реализаций $\prod_{i=1}^{z} l_i$, где l_i - число различных значений признака X_i. Число признаков может быть $z = 100 - 300$, а объем выборки $N = 20 - 50$.

4. Не делается предположения о независимости признаков, т.е. допускается сильная корреляция между признаками. В том случае, когда исходная система признаков разбита на независимые группы, а в группах выполнено условие $N \geq \mathcal{B}$ (где $\mathcal{B}$ -число возможных реализаций в группе), то в работе [6] предложено наилучшее решающее правило. В том случае, когда по каким-либо причинам систему признаков не удается разбить на независимые группы, предлагается подход, изложенный ниже.

§ 2. О выборе эффективного решающего правила и об оценке его качества

Пусть дана система признаков $X_1,\dots,X_z$, обучающая выборка объема N. Признак X_i принимает значения $x_{i1},\dots,x_{il_i}$. Для простоты рассматривается два образа.

первому образу соответствует распределение $\rho(x,I)=\rho(I)\cdot\rho(x/I)$, второму образу $\rho(x,\bar{I})=\rho(\bar{I})\cdot\rho(x/\bar{I})$, где $x=(x_1,\dots,x_z)$.

Рассмотрим множество возможных решающих правил $\{F_\theta\}\,\theta\in\Xi$, построенных на системе признаков $X_1,\dots,X_z$. Решающее правило F_θ есть функция f, заданная на некотором разбиении A_B^θ пространства возможных реализаций E

$$A_B^\theta=\{A_1^\theta,\dots,A_i^\theta,\dots,A_B^\theta\}$$

и принимающая значения I, II.

Дадим определение разбиения A_B^θ. Определим конъюнктивный член (к.ч.) как логическое высказывание $y_{j_1}\wedge y_{j_2}\wedge\dots\wedge y_{j_\Delta}$, где $y_j\in\{x_{j1},\dots,x_{jl_j},\bar{x}_{j1},\dots,\bar{x}_{jl_j}\}$ для $j=1,\dots,z$.

Тогда

$$A_1^\theta=k.z.$$

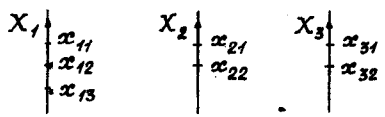
$$A_2^\theta=k.z.\wedge\bar{A}_1^\theta$$

$$\dots\dots\dots$$

$$A_B^\theta=E\wedge\bar{A}_1^\theta\wedge\dots\wedge\bar{A}_{B-1}^\theta.$$

(В дальнейшем будем опускать индекс θ там, где это не вызовет недоразумений).

Приведем пример разбиения:



$$A_1=x_{11}\wedge x_{21}$$

$$A_2=x_{11}\wedge x_{31}\wedge(\overline{x_{12}\wedge x_{21}})$$

$$A_3=x_{12}\wedge x_{22}$$

$$A_4=x_{13}\wedge x_{22}$$

$$A_5=\bar{A}_1\wedge\bar{A}_2\wedge\bar{A}_3\wedge\bar{A}_4.$$

Если известно распределение вероятностей на множестве всех допустимых реализаций в пространстве E, то на разбиении A_B

тоже задается распределение $\rho(A_i/I),\rho(A_i/\bar{I})$ для $i=1,\dots,B$. Обозначим $\rho(A_i/I)$ через $\rho(i/I)$ и $\rho(A_i/\bar{I})$ через $\rho(i/\bar{I})$.

Известно, что оптимальное решающее правило при известных вероятностях задается так

$$f(A_i)=\begin{cases}\text{I, если } \rho(I)\cdot\rho(A_i/I)\geq(1-\rho(I))\cdot\rho(A_i/\bar{I}),\\ \text{II, если } \rho(I)\cdot\rho(A_i/I)<(1-\rho(I))\cdot\rho(A_i/\bar{I}),\end{cases}\qquad(1)$$

и вероятность правильной классификации

$$\rho_{п.к}=\sum_{i=1}^{B}\max_{I,\bar{I}}[\rho\cdot\rho(A_i/I),(1-\rho)\cdot\rho(A_i/\bar{I})].$$

Но, как правило, неизвестен вектор истинных вероятностей $\rho=[\rho(I),\rho(i/I),\rho(i/\bar{I})]$ $i=1,\dots,B$, а известна лишь обучающая выборка объема N. Для каждого разбиения A_B^θ определены n_i – числа реализаций закономерности A_i^θ при условии I-го образа и m_i – число реализаций A_i^θ при условии второго образа. Вместо вектора ρ имеем оценку $\bar{\rho}=[\bar{\rho}(I),\bar{\rho}(i/I),\bar{\rho}(i/\bar{I})]$, где $\bar{\rho}(I)=\frac{n}{N}$, $(n=\sum_{i=1}^{B}n_i,\ m=\sum_{i=1}^{B}m_i)$, $\bar{\rho}(i/I)=\frac{n_i}{n}$, $\bar{\rho}(i/\bar{I})=\frac{m_i}{m}$.

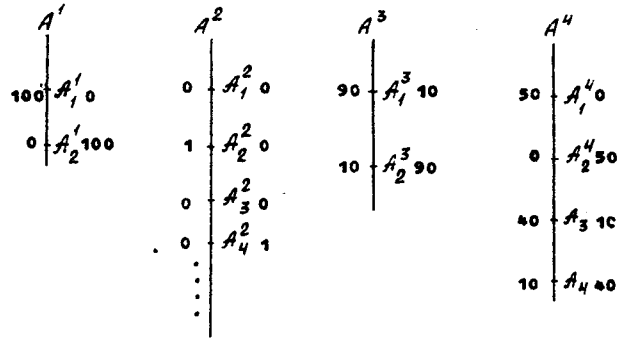
Решающее правило (I) принимает вид:

$$f(A_i)=\begin{cases}\text{I, если } n_i\geq m_i\\ \text{II, если } n_i<m_i.\end{cases}\qquad(2)$$

Этому правилу соответствует некоторая вероятность правильной классификации $\rho_{п.к.}(N,\bar{\rho})$, которую можно определить, если бы знали вектор вероятностей ρ. Вероятность $\rho_{п.к.}(N,\bar{\rho})=\xi$ является случайной величиной, так как представляет собой функцию случайных наблюдений, составляющих обучающую выборку. Эта случайная величина имеет некоторое распределение $F(\xi)$. Ясно, что ξ меньше или равна $\rho_{п.к.}$, полученной при оптимальном решающем правиле (I).

Возникает следующий вопрос. Как оценить качество выбранного разбиения для того, чтобы выбрать наилучшее при условии малой выборки? Ответу на этот вопрос и посвящена настоящая работа. Рассмотрим пример*). Пусть даны разбиения A^1, A^2, A^3, A^4:

*) Для каждого разбиения слева указаны числа n_i, справа числа m_i. Для разбиения A^2 число B велико (например, $B=10^4$), и слева 100 единиц, справа также 100 единиц, причем совпадающих единиц для обоих образов нет.



Интуитивно порядок предпочтения указанных разбиений можно указать следующий $A^1 \succ A^3 \succ A^4 \succ A^2$. Это правило предпочтения можно сформулировать так. Правило должно быть простым (заданным на малом числе закономерностей) и в то же время хорошо разделять первый и второй образ.

Попытаемся теперь получить численные статистические оценки для установления порядка предпочтения для таких разбиений. Казалось бы, что просто и естественно любые два разбиения сравнивать по следующей оценке

$$\bar{p}_{п.к.} = \sum_{i=1}^{\mathcal{B}} \max_{I,\bar{I}} [\bar{p}(I)\cdot\bar{p}(i/I),(1-\bar{p}(I))\bar{p}(i/\bar{I})],$$

где оценки для вероятностей получены по той же выборке, по которой строилось решающее правило. Но с точки зрения это оценки разбиения A^1 и A^2 эквивалентны. Неудовлетворительность этой оценки объясняется тем, что, несмотря на то, что оценки $\bar{p}(I)$, $\bar{p}(i/I)$, $\bar{p}(i/\bar{I})$ несмещенные и состоятельные, оценка $\bar{p}_{п.к.}$ для вероятности правильной классификации является смещенной. Можно привести пример, когда это смещение очень значительно. Пусть $\mathcal{B}$ очень велико ($\mathcal{B} = 10^3$), $p(I) = 1 - p(\bar{I}) = 1/2$. Распределения $p(i/I)$ и $p(i/\bar{I})$ равномерны на $A_1, \ldots, A_{\mathcal{B}}$, объем выборки $N = 50$. Вероятность $p_{п.к.} = 1/2$, и в то же время очевидно, что $M(\bar{p}_{п.к.}) \approx 1$. В этом случае используют известный эвристический прием для оценки качества решающего правила, который сводится к выделению всех реализаций поочередно в контроль. Но в этом случае показатель качества не различит разбиений A^3 и A^4 из приведенного выше примера.

Далее, возникает еще один вопрос. Пусть даны два разбиения A^1 и A^2, и известно, что $p_{п.к}(A^1) > p_{п.к}(A^2)$. Всегда ли предпочтительней разбиение A^1 по отношению к A^2 для построения решающего правила на основе выборки малого объема?

Рассмотрим пример. Для первого разбиения

$$A^1_{\mathcal{B}} = \{A^1_1, \ldots, A^1_{\mathcal{B}}\}, \quad \mathcal{B} = 10^3, \quad p = 1-p = 1/2,$$

$$p(i/I) = \begin{cases} 1/500, \\ \text{либо } 0 \end{cases}, \qquad p(i/\bar{I}) = \begin{cases} 1/500, \\ \text{либо } 0, \end{cases}$$

$$p(i/I)\cdot p(i/\bar{I}) = 0.$$

Для второго разбиения

$$A^2_{\mathcal{B}} = \{A^2_1, A^2_2\},$$

$$p(1/I) = 0.9, \quad p(2/I) = 0.1, \quad p(1/\bar{I}) = 0.1, \quad p(2/\bar{I}) = 0.9.$$

Ясно, что $p_{п.к.}(A^1_{\mathcal{B}}) = 1$, $p_{п.к.}(A^2_{\mathcal{B}}) = 0{,}9$. И в то же время для построения решающего правила по обучающей выборке малого объема (например, $N = 20$) предпочтительнее второе разбиение. Таким образом, при построении решающего правила по обучающей выборке качество разбиения не определяется $p_{п.к.}$, которую можно получить, зная истинные вероятности.

Для сравнения двух любых разбиений A^1 и A^2 должны сопоставляться соответствующие им распределения F_1 и F_2 для случайных величин ξ_1 и ξ_2 (вероятностей правильных классификаций при первом и втором разбиении). Рассмотрим это подробнее.

Если для всех ξ ($0 \le \xi \le 1$) $\mathcal{P}(\xi_1 > \xi) \ge \mathcal{P}(\xi_2 \ge \xi)$, то распределение F_1 предпочтительнее распределения F_2 . Но этот порядок будет лишь частичным на множестве возможных функций распределения. Для задания полного порядка можно предположить два подхода:

1) $A^1 \succ A^2 \Leftrightarrow M\xi_1 \ge M\xi_2$, т.е. в среднем получаем большую $p_{п.к.}$ для разбиения A^1 .

2) $A^1 \succ A^2 \Leftrightarrow \mathcal{P}(\xi_1 \ge 1-\varepsilon) \ge \mathcal{P}(\xi_2 \ge 1-\varepsilon)$, т.е. для разбиения A^1 с большей достоверностью получаем $p_{п.к.}$, превышающей некоторый порог $1-\varepsilon (0 < \varepsilon < 1)$.

Заметим, что обычно, если заданы распределения выигрыша, то используют порядок I, но это оправдано, если игра повторяется

достаточно много раз.При построении же решающего правила мы играем только один раз. Поэтому далеко не очевидно, что всегда нужно пользоваться порядком I. Однако если A^1 много лучше A^2 по порядку I, то A^1 будет лучше A^2 и по порядку 2. В данной работе решающие правила будем сравнивать по порядку I. Рассмотрим статистический аппарат, который потребуется для решения задачи.

Пусть случайная величина η имеет непрерывную функцию распределения G, $x_1,\dots,x_n$ - выборка, соответствующая величине η, а $x_{(1)},\dots,x_{(n)}$ - упорядоченная выборка.

$$B_1=[-\infty, x_{(1)}],\dots,B_{n+1}=[x_{(n)},\infty],$$

$$u_i=\rho(B_i)=G(x_{(i)})-G(x_{(i-1)})\text{ - } i\text{ - я доля.}$$

Тогда сумма любых k долей из $n+1$ имеет бета-распределение $B_\ell(k,n-k+1)$ [7]. Для любого отрезка $[\alpha,b]$ можно указать блоки $B_{i-1},B_i,\dots,B_k,B_{k+1}$ такие, что

$$\bigcup_{j=i}^{k} B_j \subset [\alpha,b] \subset \bigcup_{j=i-1}^{k+1} B_j .$$

Откуда следует, что $\eta_1\le\rho[\alpha,b]\le\eta_2$,
где

$$\rho[\alpha,b]=F(b)-F(\alpha),$$
$$\eta_1=u_i+\dots+u_k,$$
$$\eta_2=u_{i-1}+u_i+\dots+u_k+u_{k+1}.$$

Следовательно,можно определить маргинальное распределение η_1 и η_2 и их совместное распределение, что позволит оценить $\rho[\alpha,b]$.

Аналогичный результат был получен Тьюки [8] для дискретных распределений. Приведем краткое обоснование оценок для дискретных множеств.Пусть на событиях $C_1,\dots,C_m$ $(C_i\cap C_j=\Lambda,\bigcup_{i=1}^{m}C_i=E)$ задано распределение $\{p_1,\dots,p_m\}$. Задана выборка $n(C_1)=n_1,\dots,n(C_m)=n_m$, соответствующая этому распределению $(\sum_{i=1}^{m}n_i=n)$. Для любого множества $\{C_{i_1},\dots,C_{i_k}\}=C$ необходимо оценить $\rho(C)$.

ЛЕММА.

$$\eta_1\le\rho(C_i,n(C_i)=n_i)\le\eta_2,$$

где η_1 и η_2 - случайные величины, определенные выше.

ДОКАЗАТЕЛЬСТВО. Зададим на отрезке [0,I] случайную величину следующим образом. Производим испытание с распределением $\{p_1,\dots,p_m\}$ и, если происходит событие C_i, выбираем значение η на отрезке $[\sum_{k=0}^{i-1}p_k,\ \sum_{k=0}^{i}p_k]$ $(p_0=0)$ в соответствии с равномерным распределением. Случайная величина η имеет равномерное распределение. Если сделать выборку объема η, то этой выборке $n(C_1)=n_1,\dots,n(C_m)=n_m$ будет соответствовать числовая выборка $x_1,\dots,x_n$ и ей упорядоченная выборка $x_{(1)},\dots,x_{(n)}$.

Событие C_i эквивалентно $\{\sum_{k=0}^{i-1}p_k\le\eta\le\sum_{k=0}^{i}p_k\}$, и по приведенному выше утверждению

$$\bigcup_{j=f_1}^{f_2} B_j \subset C_i \subset \bigcup_{j=f_1-1}^{f_2+1} B_j .$$

Откуда и следует требуемое утверждение.

Пользуясь вышеприведенными замечаниями,оценим предпочтительность одного разбиения перед другим.Пусть разбиение $\{A_1,\dots,A_B\}$ задано. Для установления качества этого разбиения применим лемму, доказанную ранее, к системе событий

$$\{(A_1,I),(A_1,II),\dots,(A_B,I),(A_B,II)\}.$$

По решающему правилу (2) для каждой пары $\{(A_i,I),(A_i,II)\}$ выделяем то событие, для которого число реализаций на обучающей выборке больше. Обозначим его через B_i, а второе в паре событие через $\bar{B}_i$. Тогда

$$p_{n.k}=p(\bigcup_{i=1}^{B}B_i).$$

Из леммы следует, что

$$\eta_1\le p(\bigcup_{i=1}^{B}B_i,\ n(B_i)=\max(n_i,m_i))\le\eta_2,$$

где η_1 есть объединение μ долей.
Величина

$$\mu=\sum_{i=1}^{B}\varphi_i, \qquad (3)$$

где

$$\varphi_i = \begin{cases} max(m_i, n_i)-1 \text{ при } max(m_i, n_i) \geq 1, \\ 0 \quad \text{при } max(m_i, n_i) < 1. \end{cases}$$

Величина η_1 распределена по закону $Be(\mu, N-\mu+1)$. η_2 есть объединение $\mu+2\mathcal{B}$ долей (некоторые из этих долей могут входить повторно). Для установления предпочтения используем нижнюю оценку η_1. Для упорядочения будем использовать порядок I, т.е. по величине математического ожидания.

Тогда показателем качества разбиения $A_{\mathcal{B}}^{\theta}$ будет

$$\Delta(A_{\mathcal{B}}^{\theta}) = \frac{\mu}{N+1}.$$

Рассмотрим порядок, устанавливаемый этим показателем для примера (стр. I02).

$$\Delta(A^1) = \frac{198}{201}, \quad \Delta(A^2) = 0, \quad \Delta(A^3) = \frac{178}{201}, \quad \Delta(A^4) = \frac{176}{201},$$

откуда

$$A^1 \succ A^3 \succ A^4 \succ A^2.$$

Этот статистический критерий качества позволяет выделить систему конъюнктивных членов для построения решающего правила на основе малой обучающей выборки при большом числе признаков. Заметим, что он учитывает как разделяющую способность системы, так и надежность решающего правила для нее.

Литература

I. ЛБОВ Г.С. О представительности выборки при выборе эффективной системы признаков. "Вычислительные системы", Новосибирск, 1966, вып. 22.

2. РАУДИС Ш.Ю. Оценка объема обучающей выборки классификаторов опознающих устройств. Диссертация. Каунас, 1969.

3. ВАПНИК В.Н., ЧЕРВОНЕНКИС А.Я. О равномерной сходимости частот появления событий к их вероятностям.-"Теория вероятностей и её применения", 1971, т. XVI, вып. 2.

4. ВАПНИК В.Н. Задача обучения распознаванию образов. М., Изд-во "Знание", 1971.

5. ЛБОВ Г.С., КОТЮКОВ В.И., МАНОХИН А.Н. Об одном алгоритме распознавания в пространстве разнотипных признаков. Настоящий сборник, с.108-110.

6. БОРОВКОВ А.А. О задаче распознавания образов. -"Теория вероятностей и её применения", 1971, т. XVI, вып. I.

7. УИЛКС. Математическая статистика . М., "Наука", 1967.

8. TUKEY I.W. Nonparametric estimation. III. Statistically equivalent blocks and multivariate tolerance regions - the discontinuons case. - "Ann.Math.stat.",1948,vol.12, p.30-39.

9. TUKEY. Nonparametric estimation. III. Statistically equivalent blocks and tolerance regions - the continuous case.- "Ann.Math.stat.",1947,vol.18, p.529-539.

Поступила в ред.-изд.отд.
28 мая 1973 года