ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ

Сборник трудов

Института математики СО АН СССР

1967 г. Выпуск 28

# ОБ ЭФФЕКТЕ ЗАВИСИМОСТИ МЕЖДУ ОЦЕНКАМИ КАЧЕСТВА ПРИЗНАКОВ ОПОЗНАЮЩИХ СИСТЕМ

**А.Ю. Мотуза**
**(Вильнюс)**

В работе исследуется влияние зависимости между оценками качества систем признаков и вероятностью правильных решений при выборе наилучшей из указанных систем.

При синтезе опознающих систем приходится решать так называемую задачу выбора качественных признаков. Формализуем эту задачу следующим образом. Пусть имеется множество $S = \{S_1, \dots, S_U\}$ ($U$ -объем множества $S$) систем признаков (СП). Из этого множества требуется выбрать лучшую по заранее заданному критерию (например, по наибольшей вероятности вер - ного опознавания речевых или зрительных образов) СП. Показатели качества (ПК) $Q_1, \dots, Q_U$ этих СП точно неизвестны. О соотношениях этих ПК исследователь принимает решение $\tau$ ($\tau \in R$; $R$ - заданное множество решений) по состоятельным оценкам $t_1, \dots, t_U$ монотонных (скажем возрастающих) функций величин $Q_1, \dots, Q_U$. Для подсчета каждой оценки $t_i$ ($i=1,\dots,U$) используется выборка конечного объема $N^i$ ($i=1,\dots,U$) реализаций $\bar{\xi}^i_{\nu_i}$ ($\nu_i = 1, \dots, N^i$) опознаваемых объектов (слов, геометрических фигур). Ввиду конечности объемов $N^i$ и случайного характера объектов $\bar{\xi}^i_{\nu_i}$ оценки $t_1, \dots, t_U$ случайные величины, разброс которых возрастает с убыванием $N^i$. Это приводит к ошибкам при принятии решения $\tau$. В большинстве практических случаев требуется, чтобы вероятность правильного решения $P$ была не ниже заданной величины $P^*$. Величина $P$ возрастает с увеличением объемов выборок $N^i$ ($i=1,\dots,U$). Но увеличение этих объемов связано с затратами. В тех случаях,

когда оценки $t_1,\dots,t_u$ независимы, проблема определения наименьшего объема выборки $N^1_{min}=\dots=N^u_{min}=N_{min}$, при котором вероятность ошибочного решения $1-P$ не превышает заданной величины $1-P^*$, исследована в ряде работ[I-5]. Причем существуют практические методы решения этой задачи. Статистическую независимость оценок $t$ можно достичь рандомизацией выбора реализаций $\bar{\xi}$. При этом для вычисления каждой из оценок $t_i$ $(i=1,\dots,u)$ должны быть использованы различные выборки. Такой прием приводит к сильному завышению общего числа $N^1+\dots+N^u$ реализаций опознаваемых объектов $\bar{\xi}$. Во многих практических случаях все оценки $t_1,\dots,t_u$ подсчитываются по одной выборке объема $N$ опознаваемых объектов

$\bar{\xi}^1_{v_i}=\dots=\bar{\xi}^u_{v_i}=\bar{\xi}_v$ $(v=1,\dots,N)$. В этом случае величины $t_1,\dots,t_u$ статистически зависимы.

В настоящей работе исследуем, как влияет зависимость между оценками $t_1,\dots,t_u$ на вероятность правильного решения $P$. Исследования проведем в случае гауссовского распределения величин $t_1,\dots,t_u$. Множество решений $R$ определим следующим образом $R=\{r_1,\dots,r_i,\dots,r_u\}$, где $r$ - такое решение, что между ПК $Q_1,\dots,Q_u$ существует следующее соотношение:

$$Q_i \geq \max(Q_1,\dots,Q_u)-\delta. \qquad (1)$$

Здесь $\delta$ - так называемый коэффициент неточности [5].

В этом случае решение $r_i$ будет следующей функцией оценок $t_1,\dots,t_u$:

$$r_i=\begin{cases} 1, \text{ если } t_i=\max(t_1,\dots,t_u); & (2)\\ 0 - \text{в другом случае.}\end{cases}$$

Вероятность правильного решения $P$ тогда определяется по формулам

$$P=\sum_{i=1}^{u} P_i,$$

где

$$P_i=\begin{cases} P_i(r_i)=P(t_i=\max(t_1,\dots,t_u)) & (3)\\ \text{при } Q_i>\max(Q_1,\dots,Q_u);\\ 0 - \text{в других случаях.}\end{cases}$$



Вероятность $P_i(r_i)$ можно определить, зная условные законы плотности ) распределения вероятностей случайных величин $t_1,\dots,t_u$. Пусть эти величины распределены по нормальному закону и их математические ожидания равны показателям качества $Q_1,\dots,Q_u$, т.е.

$$f(\bar{t}\,|\,\bar{Q})=(2\pi)^{-u/2}|\Sigma|^{-1/2}\exp\left[-(\bar{t}-\bar{Q})'\Sigma^{-1}(\bar{t}-\bar{Q})\right], \qquad (4)$$

где

$$\Sigma=\begin{pmatrix} K_{11}&\dots&K_{1u}\\ \vdots&&\vdots\\ K_{u1}&\dots&K_{uu}\end{pmatrix},$$

$\Sigma$ - матрица коэффициентов корреляции величин $t_1,\dots,t_u$;

$$t=\begin{pmatrix} t_1\\ \vdots\\ t_u\end{pmatrix};\qquad \bar{Q}=\begin{pmatrix} Q_1\\ \vdots\\ Q_u\end{pmatrix},$$

где $t$ и $\bar{Q}$ - векторы оценок $t_i$ и ПК, $Q_i$, соответственно.

Вероятности $P_i(r_i)$, а тем самым и вероятность правильного решения $P$, можно определить, используя теорию многомерного статистического анализа нормально распределенных величин [6]. Она будет

$$P_i(r_i)=\int_{-\infty}^{\Delta Q_{i1}}\dots\int_{-\infty}^{\Delta Q_{iu}}(2\pi)^{-\frac{1}{2}(u-1)}|\rho|^{-\frac{1}{2}}\exp(-\bar{x}'\rho^{-1}\bar{x})\,dx_1\dots dx_u,$$

где $\bar{x}$ - $u-1$ - мерный случайный вектор с равным $0$ математическим ожиданием и корреляционной матрицей $\rho$;

$$\Delta Q_{ij}=(Q_i-Q_j)\cdot\left[K_{ii}+K_{jj}-2a_{ij}(K_{ii}\cdot K_{jj})^{1/2}\right]^{-1/2} \qquad (6)$$

Элементы $\rho_{kj}$ матрицы $\rho$ определяются по формуле

$$\rho_{kj}=(K_{ii}+K_{jk}-K_{ik}-K_{ij})\left[(K_{ii}+K_{kk}-2K_{ij})(K_{ii}+K_{jj}-2K_{ij})\right]^{-1/2} \quad (7)$$

$$a_{ij}=K_{ij}(K_{ii}\cdot K_{jj})^{-1/2},$$

где $a_{ij}$ - коэффициент корреляции оценок $t_i, t_j$ $(i,j=1,\ldots,U)$.

Из формулы (5) видно, что вероятность $P_i(\tau_i)$ является возрастающей функцией аргументов $\Delta Q_{ij}$ $(j=1,\ldots,U,\ i\neq j)$. Величины $\Delta Q_{ij}$, как видно из формулы (6), при $Q_i > Q_j$ являются возрастающими функциями коэффициентов корреляции $a_{ij}$. Следовательно, вероятность $P_i(\tau_i)$ возрастает с возрастанием зависимости между оценкой $i$ -го показателя качества $t_i$ и остальными оценками $t_j$ $(j=1,\ldots,U;\ i\neq j)$. Кроме того, из формул (5) и (6) видно, что при $K_{ii}=K_{jj}$ выполняется следующее равенство:

$$\lim_{\substack{a_{ij}\to 1\\ (j=1,\ldots,U)}} P_i(\tau_i)=1. \quad (8)$$

Так как вероятность правильного выбора $P$ является суммой вероятностей $P_i(\tau_i)$ (формула (2)), то она тоже возрастает с увеличением корреляции $a_{ij}$ $(i,j=1,\ldots,U;\ i\neq j)$ между оценками $t_1,\ldots,t_U$ функции показателей качества $Q_1,\ldots,Q_U$

Аналогичным путем можно показать, что вероятность $P$ является возрастающей функцией от корреляционных коэффициентов $a_{ij}$ и при других множествах $R$ решений $\tau$, встречаемых при выборе качественных признаков опознающих систем.

Известно, что в большинстве случаев, при использовании той же самой выборки опознающих объектов $\bar{\xi}_1,\ldots,\bar{\xi}_N$ для подсчета всех оценок качества $t_1,\ldots,t_U$ коэффициенты корреляций $a_{ij}>0$. В таких случаях вероятность $P$ можно заменить её оценкой $P'$, которая легко подсчитывается, допуская, что оценки $t_1,\ldots,t_U$ независимы. При этом будет выполнено неравенство $P>P'$. Следовательно, если нам надо определить такой объем выборки $N$, при котором выполняется неравенство $P>P^*$ в этом случае, мы можем воспользоваться существующими методами [1,2,8,4,5] выбора качественных систем



признаков. При этом при вычислениях вероятность $P$ будет замена её оценкой $P'$.

В заключение заметим, что полученные в настоящей работе результаты легко обобщаются при использовании предельных теорем для слабо зависимых случайных величин, для случаев биномиальных и Гамма-распределений оценок $t_1,\ldots,t_U$.

## Литература


1. Paulson E. On the comparison of several experimental categories with a control. Ann.Math.Stat., 1952, 23, 239–46.
2. Bechhofer R.E. A single sample multiple decision procedure for ranking the means of normal populations with known variances. Ann.Math.Statist., 1954, 25, 16–39.
3. Mace A.E. Sample-size determibation, 1964, New York, Reinhold Publishing Corporation.
4. Chamber M.L. and Jarrat P. Use of double sampling for selecting best population. Biometrika, 1964, 51, 49–64.
5. А.Ю.Мотуза. Алгоритмы выбора полезных признаков опознающих систем. Рефераты докладов I-го Всесоюзного симпозиума по статистическим проблемам в технической кибернетике. Москва, 1967, ч.II, 9–10.
6. Т.Андерсон. Введение в многомерный статистический анализ. ГИФМЛ, Москва.


Поступила в редакцию
I/X-1967г.