УДК 621.391

И.К. ВАСИЛЬЕВА

Национальный аэрокосмический университет им. Н.Е. Жуковского "ХАИ", Украина

О ВЛИЯНИИ КОРРЕЛЯЦИИ ПРИЗНАКОВ НА ДЛИТЕЛЬНОСТЬ ПОСЛЕДОВАТЕЛЬНОЙ ПРОЦЕДУРЫ РАСПОЗНАВАНИЯ

Выполнено исследование влияния корреляционных связей между компонентами многомерных признаков на длительность последовательной процедуры распознавания при симметричных порогах. Контрольные выборки признаков двух классов объектов были получены путем моделирования двумерных случайных величин, распределенных по нормальному закону с равными корреляционными матрицами. Коэффициенты взаимной корреляции компонент признаков варьировались в диапазоне от – 0,99 до 0,99. В качестве показателей длительности последовательного анализа определялись статистические оценки среднего и максимального объемов контрольных выборок, требующихся для обеспечения одинаковых уровней вероятностей ошибок первого и второго рода ($\alpha = \beta = 0{,}01$). Показано, что значения оценок среднего и максимального объемов выборок существенно зависят от силы и направления корреляционной взаимосвязи между компонентами признаков классов.

Ключевые слова: *распознавание, анализ Вальда, корреляционная матрица, объем контрольной выборки.*

Введение

Жесткое ограничение времени на принятие решения является одним из наиболее существенных требований, предъявляемых к интегрированным автоматизированным системам, осуществляющим непрерывный мониторинг, обнаружение и распознавание классов объектов в реальном масштабе времени. При этом повышение размерности признакового пространства может оказаться единственным средством увеличения достоверности до требуемого уровня. С практической точки зрения представляет интерес задача оптимизации суммарного количества наблюдений (определяемого объемом контрольной выборки n и размерностью вектора измерений p), необходимого для обеспечения заданного гарантированного уровня достоверности распознавания при заданном наименьшем возможном расстоянии между классами, в качестве которого из практических соображений естественно взять реальную точность измерения этого расстояния в распознающих системах. Если считать величину p заданной (равной числу каналов приема информации), то выбор величины n (объема выборки) можно осуществлять по методу последовательного анализа Вальда [1, 2].

Следует отметить, что до настоящего времени основным препятствием к широкому использованию методов последовательного анализа считается недостаточность информации о характеристиках продолжительности данной процедуры. С учетом того, что признаки, несущие информацию об объектах распознавания, как правило, являются зависимыми (коррелированными), задачей данного исследования являлось изучение влияния силы и направленности корреляционных связей компонент двумерных нормальных признаков на распределение длительности анализа Вальда при симметричных порогах.

В работах [3 – 5] показано, что использование сильно коррелированных признаков может приводить как к возрастанию достоверности байесовского распознавания, так и к ее заметному снижению. Изменения величин вероятностей ошибок связаны с трансформацией условных по классам плотностей распределения вероятности, обусловленной корреляцией компонент векторного признака (поворот и масштабирование осей эллипса рассеяния). Обозначим собственные вектора, соответствующие таким собственным значениям λ_i корреляционной матрицы, что $|\lambda_i| < 1$, через $\vec{\Phi}_i^-$, а собственные вектора, соответствующие $|\lambda_i| > 1$, – $\vec{\Phi}_i^+$. По направлениям $\vec{\Phi}_i^-$ происходит сжатие поверхности многомерной плотности распределения (снижение разброса внутри класса), вдоль направлений $\vec{\Phi}_i^+$ – растяжение данной поверхности (увеличение внутриклассового рассеяния). Если вектор разности математических ожиданий признаков имеет преимущественную ориентацию вдоль $\vec{\Phi}_i^+$, то снижение вероятности ошибочного распознавания достигается только при отрицательных значениях коэффициентов корреляции r; при $r > 0$ достоверность классического распознавания уменьшается. В противном случае с увеличением абсолютной величины коэффициента корреляции вероятность ошибки байесовского классификатора снижается на несколько порядков. Вопрос заключался в том, является ли такое объяснение достаточным и для последовательной процедуры или же коррелированные признаки в любом случае будут быстрее достигать одного из двух порогов последовательного анализа.

Для ответа на поставленный вопрос проводилось исследование влияния степени взаимной корреляции компонент двумерных признаков на статистические оценки среднего и максимального объемов контрольных выборок, требующихся для обеспечения одинаковых уровней вероятностей ошибок 1-го и 2-го рода.

1. Последовательный анализ Вальда

Последовательное правило выбора решения, в отличие от байесовского, предусматривает сравнение логарифма отношения правдоподобия

$$L(\vec{x}) = f_p(\vec{x}|a_2)/f_p(\vec{x}|a_1) \qquad (1)$$

с двумя порогами c_1 и c_2, не зависящими от априорных вероятностей наличия или отсутствия сигнала и от потерь. Нижний и верхний пороги приближенно можно выразить через заданные значения вероятностей ложной тревоги α и пропуска сигнала β:

$$c_1 = \ln[\beta/(1-\alpha)],\ c_2 = \ln[(1-\beta)/\alpha].$$

Таким образом, при последовательном анализе Вальда на каждом этапе распознавания пространство выборок наблюдений разделяют на три области: $G_1 = (-\infty, c_1]$, $G_2 = [c_2, +\infty)$ и промежуточную область $G_{пр} = (c_1, c_2)$. Если выборочное значение попадает в $G_{пр}$, то делается следующее наблюдение, и так до тех пор, пока при некотором значении n размера выборки значение $\vec{x}^*$ не попадет в одну из областей, G_1 или G_2, после чего принимается одна из гипотез – H_1: $\vec{x}^* \in a_1$ (при попадании в G_1) или H_2: $\vec{x}^* \in a_2$ (G_2).

Процедуру последовательного анализа можно рассматривать как процесс сравнения значений членов случайной последовательности логарифмов отношения правдоподобия (1) выборок возрастающего размера $\{\vec{x}^{(1)}\}$, $\{\vec{x}^{(1)}, \vec{x}^{(2)}\}$, $\{\vec{x}^{(1)}, \vec{x}^{(2)}, \ldots, \vec{x}^{(n)}\}$ с нижним c_1 и верхним c_2 порогами; при n-м наблюдении принимается гипотеза H_1, если

$$c_1 < \sum_{v=1}^{n-1} \ln L(\vec{x}^{(v)}) < c_2,\ \sum_{v=1}^{n} \ln L(\vec{x}^{(v)}) \le c_1$$

и гипотеза H_2, если

$$c_1 < \sum_{v=1}^{n-1} \ln L(\vec{x}^{(v)}) < c_2,\ \sum_{v=1}^{n} \ln L(\vec{x}^{(v)}) \ge c_2.$$

Критерием качества последовательного правила выбора решения обычно является минимум среднего значения размера выборки n, необходимой для принятия решения, при заданных значениях α и β.

Поскольку при последовательном анализе размер выборки n является случайной величиной, то даже при сравнительно малых средних значениях длительности процедуры возможны случаи недопустимо больших размеров выборки. Типичным примером компромиссного решения для распределения длительности процедуры является усеченный последовательный анализ, при котором заранее устанавливается максимальное значение объема выборки n_{max}. При достижении n_{max} последовательная процедура заканчивается и соответствующее отношение правдоподобия сравнивается не с двумя порогами, c_1 и c_2, а только с одним, $c_{ус}$, в результате чего обязательно принимается одно из решений.

2. Методика постановки эксперимента

Методология обработки данных при наличии корреляционных связей наиболее развита для многомерного нормального закона распределения и, как правило, в практических задачах распознавания эталонными описаниями объектов служат статистические модели многомерных нормальных совокупностей вида [1]:

$$f(\vec{x}) = (2\pi)^{-p/2} |\mathbf{R}|^{-1/2} \exp\left[-\frac{1}{2}(\vec{x}-\vec{m})^T \mathbf{R}^{-1} (\vec{x}-\vec{m})\right], \qquad (2)$$

где $\vec{m}$ – вектор математического ожидания (МО) для класса объектов в p-мерном пространстве **X**;

R – корреляционная матрица (КМ), элементы которой – центральные моменты второго порядка, образованные составляющими случайного вектора.

Таким образом, в качестве объектов исследования рассматривались два класса, эталонными описаниями которых являлись двумерные нормальные распределения с МО $\vec{m}_1$, $\vec{m}_2$ и с равными КМ. Дисперсии компонент признаков приняты единичными, коэффициенты взаимной корреляции r_{12} варьировались в диапазоне [– 0,99; 0,99].

Для моделирования двух некоррелированных стандартных нормальных компонент случайного вектора $\vec{x}$ использовался алгоритм, основанный на центральной предельной теореме. Преобразование вектора $\vec{x}$ в вектор $\vec{x}_k^*$, реализации которого имитировали отсчеты признака k-го класса, осуществлялось с помощью уравнения

$$\vec{x}_k^* = \mathbf{\Phi}\mathbf{\Lambda}^{1/2}\vec{x} + \vec{m}_k^*, \qquad (3)$$

где **Φ** и $\mathbf{\Lambda} = \mathrm{diag}(\lambda_i)$ – матрицы собственных векторов и собственных значений **R**, соответственно.

На рис. 1 изображены эллипсы рассеяния (по уровню 1σ) двумерных коррелированных признаков, соответствующих двум классам объектов.

В первом случае (рис. 1, а) были приняты следующие характеристики распределения:

$$\vec{m}_1^T = (0, 0),\ \vec{m}_2^T = (1, 0);\ \vec{\sigma}_1^T = \vec{\sigma}_2^T = (1, 1).$$

Для второго случая (рис. 1, б):

$$\vec{m}_1^T = (0, 0),\ \vec{m}_2^T = (1, 1);\ \vec{\sigma}_1^T = \vec{\sigma}_2^T = (1, 1).$$

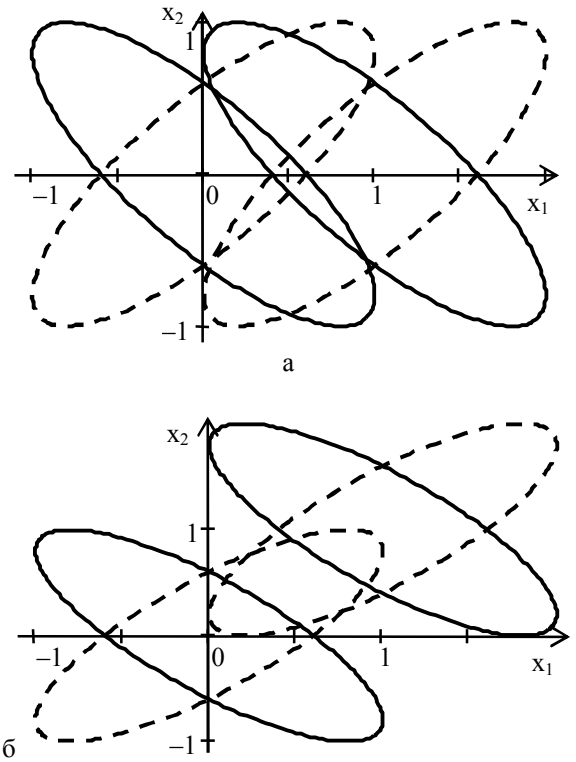

Рис. 1. Эллипсы рассеяния по уровню 0,8413 (1σ) (сплошная линия – $r_{12} = -0{,}8$, штриховая – $r_{12} = 0{,}8$) двумерных нормальных признаков с $\vec{\sigma}^{T} = (1,\ 1)$: а – $\vec{m}_1^{T} = (0,\ 0), \vec{m}_2^{T} = (1,\ 0)$, б – $\vec{m}_1^{T} = (0,\ 0), \vec{m}_2^{T} = (1,\ 0)$

Очевидно, что ошибка байесовского классификатора в первом случае будет тем ниже, чем больше абсолютное значение r_{12}. Во втором случае на вероятность ошибки будет влиять знак коэффициента корреляции: различимость коррелированных признаков лучше, если $r_{12} < 0$ и хуже, если $r_{12} > 0$.

Для исследования влияния корреляции на длительность последовательной процедуры распознавания генерировались выборки с указанными выше параметрами (с объемом генеральной совокупности N = 500). Согласно методике анализа Вальда оценивались необходимые объемы выборок каждого класса $n^{(1)}$ и $n^{(2)}$ при равных вероятностях ложной тревоги и пропуска сигнала, $\alpha = \beta = 0{,}01$. Появление классов считалось априори равновероятным, т.о., общий требуемый объем n рассчитывался как среднее $n^{(1)}$ и $n^{(2)}$. Количество актов распознавания классов a_1 и a_2 представляло реализации случайной целочисленной величины M, соответствующей условию

$$N - n^{*}_{max} < \sum_{i=1}^{M} n_i^{(k)} \le N,$$

где n^{*}_{max} – предварительная оценка сверху максимального объема контрольной выборки, $n^{*}_{max} = 50$; $n_i^{(k)}$ – i-я реализация случайной величины n для k-го класса, k = 1, 2.

По полученным реализациям случайных величин $n^{(1)}$ и $n^{(2)}$ были построены эмпирические распределения величины общего объема выборки n, необходимого для обеспечения заданных уровней α и β и определены статистические оценки минимального n_{min}, максимального n_{max} и среднего n_{cp} значений n.

3. Анализ зависимости объема выборки от степени корреляции признаков

Эксперимент по последовательному распознаванию классов a_1 и a_2 с общей корреляционной матрицей **R** проводился для 21 значения коэффициента корреляции r_{12} компонент двумерных признаков классов; предельные значения r_{12} составляли ± 0,99, остальные, начиная с – 0,9, изменялись с шагом 0,1.

Исследование влияния r_{12} на характеристики длительности последовательной процедуры распознавания проводилось для двух случаев взаимной ориентации векторов $\left(\vec{m}_1 - \vec{m}_2\right)$ и $\vec{\Phi}_i^{+}$.

Первый случай: $\vec{m}_1^{T} = (0,\ 0)$, $\vec{m}_2^{T} = (1,\ 0)$ (см. рис. 1, а). При этом угол θ между векторами $\left(\vec{m}_1 - \vec{m}_2\right)$ и $\vec{\Phi}_i^{+}$ составляет ± 45°:

$$\left\langle \left(\vec{m}_1 - \vec{m}_2\right), \vec{\Phi}^{+} \right\rangle \Big/ \left\| \vec{m}_1 - \vec{m}_2 \right\| \left\| \vec{\Phi}^{+} \right\| = \pm\sqrt{2}/2.$$

Второй случай: $\vec{m}_1^{T} = (0,\ 0)$, $\vec{m}_2^{T} = (1,\ 1)$ (рис. 1, б). В зависимости от знака r_{12} угол θ принимает значения либо 45° (т.е. ориентация вектора разности МО не совпадает с направлениями собственных векторов **R**), либо 0° (т.е. вектор разности МО коллинеарен собственному вектору, соответствующему большему собственному числу **R**):

$$\frac{\left\langle \left(\vec{m}_1 - \vec{m}_2\right), \vec{\Phi}^{+} \right\rangle}{\left\| \vec{m}_1 - \vec{m}_2 \right\| \left\| \vec{\Phi}^{+} \right\|} = \begin{cases} 1, & \text{åñëè } r_{12} > 0; \\ \sqrt{2}/2, & \text{åñëè } r_{12} < 0. \end{cases}$$

Гистограммы общего объема контрольной выборки n для ряда фиксированных значений r_{12} показаны на рис. 2, 3. Для первого исследуемого случая длительность последовательного распознавания, оцениваемая величиной n, заметно снижается при увеличении абсолютной величины коэффициента корреляции r_{12}; при $r_{12} = \pm 0{,}99$ $n_{max} = 1$. Во втором случае при $r_{12} < 0$ угол θ, как и в предыдущем исследовании, равен 45°, и увеличение абсолютного значения r_{12} приводит к уменьшению всех характеристик n (при $r_{12} = -0{,}99$ $n_{max} = 1$). Для $r_{12} > 0$ наблюдается обратный эффект: с увеличением r_{12} оценки характеристик n возрастают.

На рис. 4, 5 изображены виды зависимостей максимального n_{max} и среднего n_{cp} значений объема выборки от величины коэффициента корреляции r_{12}.

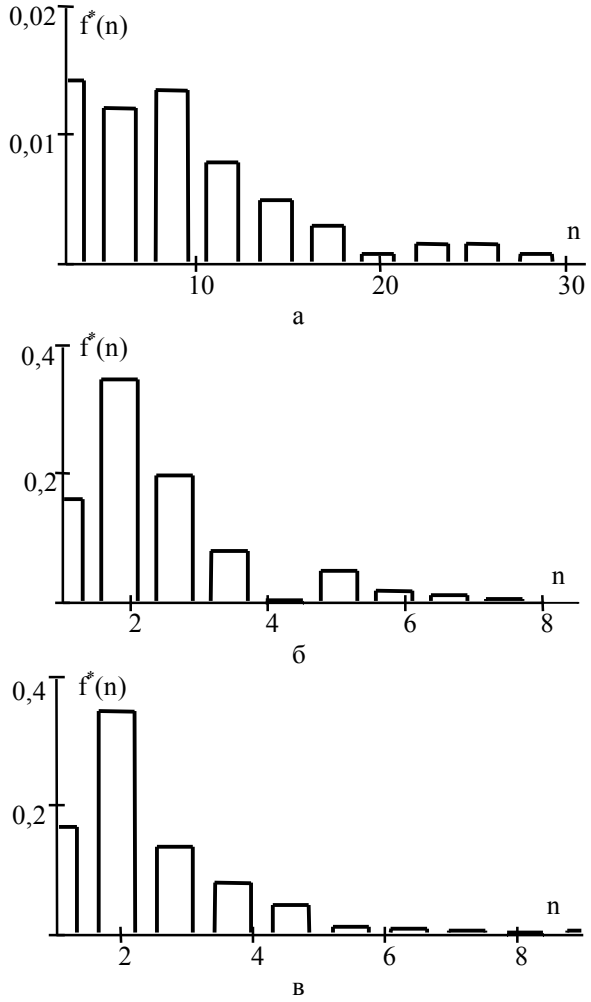

Рис. 2. Гистограммы эмпирического распределения случайного объема контрольной выборки n для классов a_1 и a_2 (первый случай, $\theta = \pm 45°$) при разных значениях коэффициента корреляции:
а – $r_{12} = 0$ ($n_{max} = 31$, $n_{min} = 3$, $n_{cp} = 10{,}6$);
б – $r_{12} = 0{,}9$ ($n_{max} = 9$, $n_{min} = 1$, $n_{cp} = 2{,}5$);
в – $r_{12} = -0{,}9$ ($n_{max} = 9$, $n_{min} = 1$, $n_{cp} = 2{,}5$)

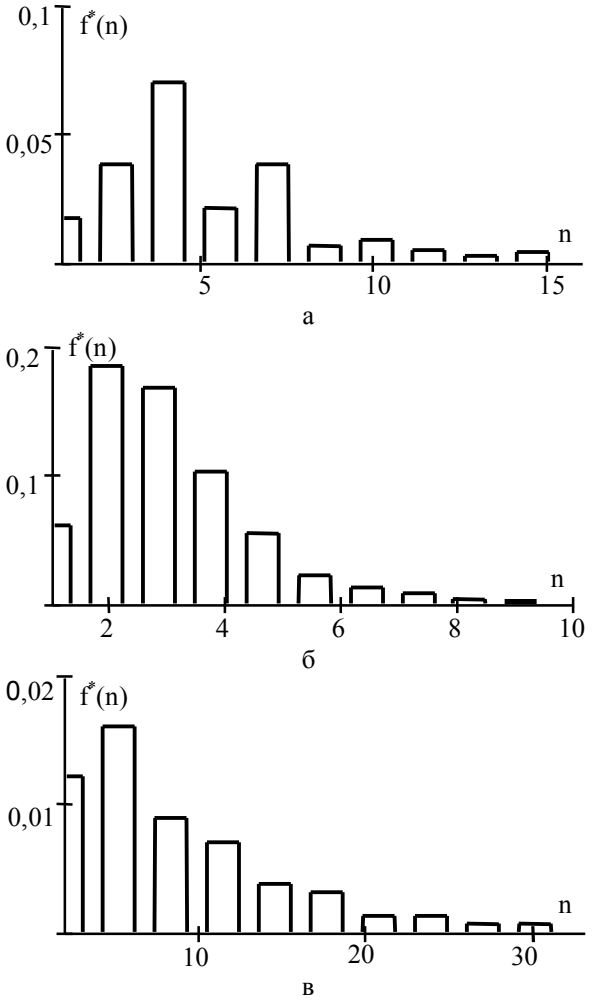

Рис. 3. Гистограммы эмпирического распределения случайного объема контрольной выборки n для классов a_1 и a_2 (второй случай, $\theta \in \{45°, 0°\}$) при разных значениях коэффициента корреляции:
а – $r_{12} = 0$ ($n_{max} = 17$, $n_{min} = 1$, $n_{cp} = 6$);
б – $r_{12} = -0{,}5$ ($n_{max} = 9$, $n_{min} = 1$, $n_{cp} = 3$);
в – $r_{12} = 0{,}99$ ($n_{max} = 32$, $n_{min} = 4$, $n_{cp} = 11$)

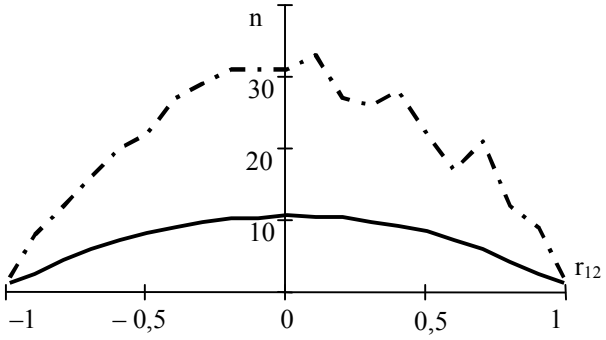

Рис. 4. Зависимость статистических оценок характеристик случайного объема выборки n от величины коэффициента корреляции r_{12} (сплошная линия – n_{cp}, штрихпунктирная – n_{max}) при $\theta = \pm 45°$

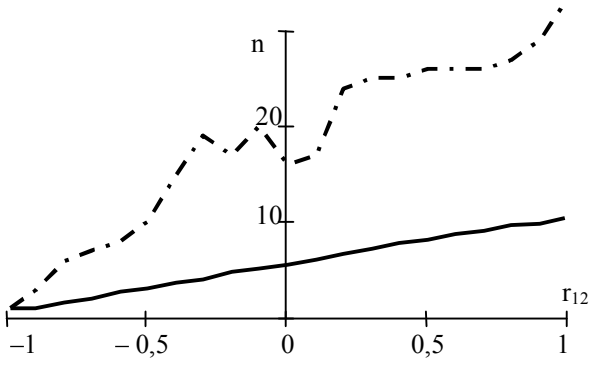

Рис. 5. Зависимость статистических оценок характеристик случайного объема выборки n от величины коэффициента корреляции r_{12} (сплошная линия – n_{cp}, штрихпунктирная – n_{max}) при $\theta \in \{45°, 0°\}$

Заключение

Результаты исследования влияния корреляции компонент двумерных нормальных признаков с равными корреляционными матрицами на продолжительность последовательного распознавания согласуются с результатами байесовской классификации, представленными в [5], и позволяют сделать следующие выводы. Изменение в ту либо другую сторону критерия качества классификатора (для анализа Вальда – среднего объема контрольной выборки, необходимой для обеспечения гарантируемых уровней вероятности ошибок 1-го и 2-го рода; для классического распознавания – вероятности ошибки) связано с взаимной ориентацией вектора разности МО признаков и собственного вектора КМ, соответствующего большему собственному числу. Если указанные вектора коллинеарны и $r_{12} < 0$, или ортогональны и $r_{12} > 0$, или если угол между векторами близок к 45°, то усиление корреляционной взаимосвязи компонент признаков (т.е., увеличение абсолютной величины r_{12}) приводит к существенному увеличению критерия качества распознавания.

Литература

1. Фомин Я.А. Статистическая теория распознавания образов / Я.А. Фомин, Г.Р. Тарловский. – М.: Радио и связь, 1986. – 264 с.

2. Храбростин Б.В. Синтез и анализ оптимальных решающих правил селекции целей в облаке случайных рассеивателей при полном поляризационном зондировании пространства / Б.В. Храбростин, М.М. Сапов, Д.Б. Храбростин // Научные ведомости БелГУ. Сер. Физика. Математика. – 2008. – Вып. 14, №9 (49). – С. 205-222.

3. Yun Fu. Correlation metric for generalized feature extraction / Fu Yun, Yan Shuicheng, T.S. Huang // IEEE Trans. on Pattern Analysis and Machine Intelligence. – 2008. – Vol. 30, Is. 12. –P. 2229-2235.

4. Васильева И.К. Об информативности коррелированных признаков объектов распознавания / И.К. Васильева, А.В. Попов // Радіоелектронні і комп'ютерні системи. – 2008. – № 3 (30). – С. 56-61.

5. Васильева И.К. Влияние степени корреляции признаков на результаты распознавания объектов по данным моделирования двумерных нормальных совокупностей / И.К. Васильева, Е.А. Панкратова // Радіоелектронні і комп'ютерні системи. – 2009. – № 1 (35). – С. 73-76.

Поступила в редакцию 7.12.2009

Рецензент: д-р техн. наук, проф., проф. кафедры производства радиоэлектронных систем Г.Я. Красовский, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина.

ПРО ВПЛИВ КОРЕЛЯЦІЇ ОЗНАК НА ТРИВАЛІСТЬ ПОСЛІДОВНОЇ ПРОЦЕДУРИ РОЗПІЗНАВАННЯ

І.К. Васильєва

Досліджено вплив кореляційних зв'язків між компонентами багатовимірних ознак на тривалість послідовної процедури розпізнавання при симетричних порогах. Контрольні вибірки ознак двох класів об'єктів отримано моделюванням двовимірних випадкових величин, розподілених за нормальним законом з рівними кореляційними матрицями. Коефіцієнти кореляції компонент ознак варіювались у діапазоні від – 0,99 до 0,99. У якості показників тривалості послідовного аналізу визначались статистичні оцінки середнього та максимального об'ємів контрольних вибірок, необхідних щоб забезпечити однакові рівні імовірностей похибок першого і другого роду ($\alpha = \beta = 0{,}01$). Показано, що значення оцінок середнього і максимального об'ємів вибірок істотно залежать від сили і напрямку кореляційного взаємозв'язку між компонентами ознак.

Ключові слова: розпізнавання, аналіз Вальда, кореляційна матриця, контрольної об'єм вибірки.

A SIGNATURE CORRELATION EFFECT ON SEQUENTIAL IDENTIFICATION DURATION

I.K. Vasilyeva

The research of the correlation relations between the components of multidimensional signature effect on sequential recognition procedure duration under symmetrical thresholds was made. The control samples of signatures of two classes of objects by modelling two-dimensional random variables distributed under the normal law with equal correlation matrix were obtained. A correlation coefficient between signature components was changed within the range of – 0,99 to 0,99. As the factors of sequential procedure analysis duration statistical estimations of maximum and average sample numbers were defined, which are required for equal first-kind and second-kind probability of error ($\alpha = \beta = 0{,}01$). It is shown that the magnitudes of maximum and average sample number estimations essentially depend on power and direction of signature correlation.

Key words: recognition, Wald analysis, correlation matrix, checks sample number.

Васильева Ирина Карловна – канд. техн. наук, доцент кафедры производства радиоэлектронных систем, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина.