УДК 681.518:004.093.1

ОЦІНКА СТАТИСТИЧНИХ СТІЙКОСТІ ТА ОДНОРІДНОСТІ НАВЧАЛЬНОЇ ВИБІРКИ ПРИ ФАКТОРНОМУ КЛАСИФІКАЦІЙНОМУ АНАЛІЗІ

А.С. Довбиш, д-р техн. наук, проф.;
М.В. Козинець, асист.;
С.Н. Котенко, студент
Сумський державний університет

Пропонується метод класифікаційного факторного аналізу в рамках інформаційно-екстремальної інтелектуальної технології. Метод дозволяє формувати в процесі керування слабо формалізованим процесом навчальної вибірки нових класів, які характеризуються статистичною стійкістю та однорідністю.

ВСТУП

Основними недоліками відомих методів кластер-аналізу [1-3], які викликають ускладнення при їх застосуванні для розв'язання практичних задач контролю та керування слабо формалізованими процесами, є:

– ігнорування перетину класів розпізнавання, що обумовлено нечіткими даними;

– наявність достатньо великого обсягу навчальної вибірки;

– модельність задач автоматичної класифікації, що обумовлює необхідність проведення попередньої нормалізації образів.

Задачі кластер-аналізу будемо поділяти на такі основні типи:

– факторний класифікаційний аналіз (ФКА) – формування нового класу розпізнавання за умови, що побудовано оптимальне розбиття для апріорного алфавіту класів при незмінних структурі та потужності словника ознак;

– кластер-аналіз – формування алфавіту класів та побудова оптимального розбиття при незмінних структурі та потужності словника ознак;

– самонавчання – формування алфавіту класів та побудова оптимального розбиття при оптимізації параметрів словника ознак.

Одним із перспективних напрямків аналізу і синтезу адаптивних систем керування є розроблення методів аналізу і синтезу адаптивних систем керування (СК) у рамках інформаційно-екстремальної інтелектуальної технології (ІЕІТ) [4], яка ґрунтується на максимізації інформаційної спроможності системи шляхом оптимізації просторово-часових параметрів функціонування за умов апріорної невизначеності, інформаційних і ресурсних обмежень. Необхідною умовою застосування на виробництві розроблених за ІЕІТ інтелектуальних систем керування, що реалізують алгоритми кластер-аналізу, є забезпечення статистичної стійкості та однорідності навчальної вибірки.

Метою статті є розроблення методу оцінки статистично стійких та однорідних навчальних вибірок нових класів, які є множиною випадкових значень ознак розпізнавання функціональних станів керованого процесу, що дозволить застосовувати алгоритмами ФКА у рамках ІЄІТ у задачах класифікаційного керування.

ПОСТАВЛЕННЯ ЗАВДАННЯ

Нехай подано $\{X_m^o \mid m = \overline{1, M}\}^{\Lambda}$ – відкритий алфавіт класів розпізнавання, який у процесі функціонування СК, що навчається, змінює свою потужність і відкриту навчальну матрицю, де N, n – кількість ознак розпізнавання та випробувань відповідно; Λ – символ відкритості множини. На етапі навчання СК за апріорно класифікованими реалізаціями образів у рамках МФСВ побудовано оптимальне в інформаційному розумінні чітке розбиття $\Re^{|M|}$ дискретного простору ознак Ω_B на M класів розпізнавання. Необхідно на етапі екзамену за алгоритмом ФКА для нового класу X_{M+1}^o сформувати статистично стійку та однорідну навчальну матрицю $\{x_{M+1}^{(j)} \mid j = \overline{1, n}\} \in \| x_m^{(j)} \mid m = \overline{1, M+1} \|^{\Lambda}$ за таким предикатним виразом:

$$(\forall x^{(j)} \in \tilde{\Re}^{\Lambda})[if\ x^{(j)} \notin \{X_m^o\}\ then\ x^{(j)} \in X_{M+1}^o],$$

де $x^{(j)}$ – двійкова реалізація-вектор образу, що розпізнається, та здійснити донавчання системи за МФСВ так, щоб максимізувати усереднене значення інформаційного КФЕ навчання СК:

$$\overline{E}^* = \frac{1}{M+1} \sum_{m=1}^{M+1} \max_{\{d\}} E_m , \qquad (1)$$

де $\{d\}$ – множина кроків навчання; E_m – інформаційний КФЕ навчання СК розпізнавати реалізації класу X_m^o.

Оскільки достовірність керуючих рішень залежить від забезпечення статистичних стійкості та однорідності багатовимірної навчальної матриці, то важливим завданням дослідження є розроблення ієрархічного алгоритму оцінки статистичних властивостей навчальних вибірок за умов їх нормальності у динамічному режимі зміни функціональних станів технологічного процесу. Умова нормальності таких вибірок на практиці забезпечує обґрунтування гіпотези компактності реалізацій образу (чіткої або нечіткої), що має місце в практичних завданнях контролю та керування.

МАТЕМАТИЧНА МОДЕЛЬ

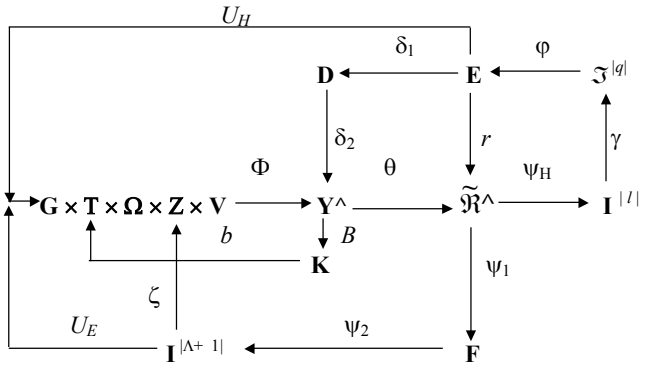
Розглянемо в рамках ІЕІТ розв'язання задачі ФКА з метою побудови в просторі ознак розпізнавання представництв, що є афінними різноманіттями, навколо яких агрегатуються нові класи. Необхідною та достатньою умовою реалізації ФКА за МФСВ є виконання нерівності $\overline{\mu}_m < c$, де $\overline{\mu}_m = \frac{1}{\mathrm{n}} \sum_{j=1}^{n} \mu_{\mathrm{m,j}}$ – усереднена функція належності; c – порогове значення, що обумовлює прийняття гіпотези $\gamma_{\Lambda+1} \in I^{|\Lambda+2|}$ – відмова від класифікації j-ї реалізації образу. Тут $I^{|\Lambda+2|}$ – множина гіпотез для відкритої абетки, де $\gamma_{\Lambda+2}$ – гіпотеза, що дозволяє формування навчальної матриці нового класу X_{Λ}^o і відповідно донавчання системи.

Нехай вхідний математичний опис СППР в режимі ФКА має таку структуру:

$$\Delta_B = < G,\ T, \Omega,\ Z,\ Y,\ K;\ \Phi,\ B >,$$

де G – простір вхідних сигналів (факторів), які діють на СППР; T – множина моментів часу зняття інформації; Ω – простір ознак розпізнавання; Z – простір можливих станів СППР; V – множина вирішальних правил; Y – вибіркова множина (вхідна навчальна матриця $\| y_{mi}^{(j)} \|$); K – ієрархічна терм-множина статистичних оцінок навчальної матриці; Φ: $G \times T \times \Omega \times Z \times V \to Y^{\Lambda}$ – оператор формування вибіркової множини Y на вході СК; B – оператор формування множини K.

Математичну модель ФКА за МФСВ подамо у вигляді категорійної моделі – діаграми відображень множин :



(2)

Контур $\Phi \to B \to b$ за результатами розвідувального аналізу забезпечує блокування алгоритму навчання СК у випадку, якщо не забезпечуються статистичні стійкість та однорідність навчальних вибірок. Оператор θ формує апріорне розбиття $\tilde{\Re}^{|\Lambda|}$, яке у загальному випадку може бути нечітким, а оператор $\Psi_H : \tilde{\Re}^{|\Lambda|} \to I^{|l|}$, де $I^{|l|} = \{\gamma_1, \ldots, \gamma_l\}$ – множина статистичних гіпотез, перевіряє основну статистичну гіпотезу $\gamma_1 : x_m^{(j)} \in X_m^o$. Після формування терм-множина $\Im^{|q|}$ точних характеристик навчання, де $q = l^2$, оператор φ обчислює значення інформаційного КФЕ (терм-множина E). Контури операторів

$\Psi_H \to \gamma \to \varphi \to r$, який оптимізує геометричні параметри розбиття

$\tilde{\Re}^{|\Lambda|}$, і $\theta \to \Psi_H \to \gamma \to \varphi \to \delta_1 \to \delta_2$, який оптимізує систему контрольних допусків D, реалізують алгоритм навчання за базовим методом IEIT– методом функціонально-статистичних випробувань МФСВ [4]. Оператор U_H оптимізує параметри плану навчання. Оператори Ψ_1, Ψ_2 і U_E реалізують алгоритм екзамену, який при кластер-аналізі функціонує паралельно з алгоритмом навчання. Тут оператор Ψ_1 обчислює терм-множину функцій належності для заданого типу класифікатора, оператор Ψ_2 здійснює дефазіфікацію, а оператор U_E

регламентує процес екзамену. Оператор ς формує додаткову навчальну матрицю і дає дозвіл на донавчання системи.

Перевагою категорійних моделей у вигляді діаграм відображень множин типу (2) є те, що вони дозволяють на етапі системного аналізу не тільки встановлювати відношення між елементами інформаційного забезпечення та інформаційними потоками оброблення інформації, але і значно полегшують розроблення структур алгоритмів різних режимів функціонування СК, що навчається.

КРИТЕРІЇ ОЦІНКИ ТА ОПТИМІЗАЦІЇ

Ієрархічна структура критеріїв оцінки статистичних стійкості та однорідності навчальної вибірки [5-7] наведено на рис. 1.

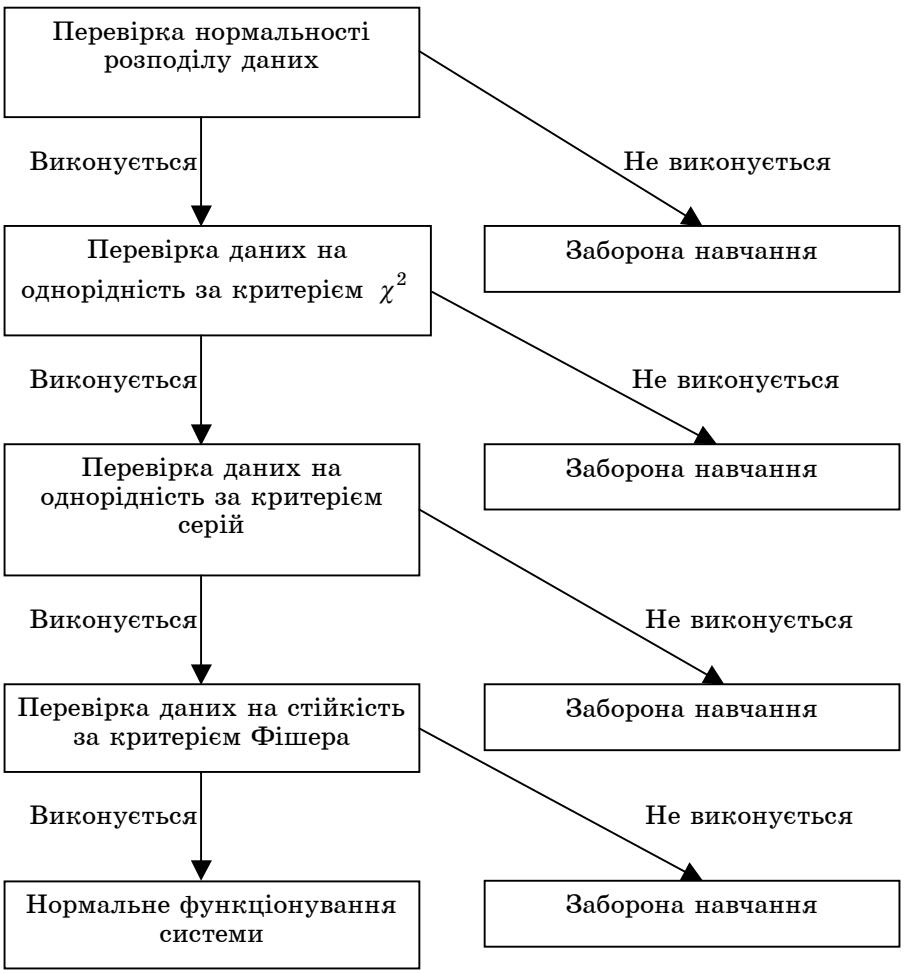


Рисунок 1 – Ієрархічна структура критеріїв оцінки статистичних стійкості та однорідності навчальної вибірки

Аналіз рис.1 показує, що невиконання основної гіпотези навіть за одним будь-яким критерієм призводить до заблокування процесу формування додаткової навчальної матриці $\| x_{\Lambda}^{(j)} \|$, а тим самим і

донавчання СК. У загальному випадку блоки навчання та екзамену припиняють функціонування за таких причин:

– перехідний процес, що відбувається при зміні функціональних станів технологічного процесу під впливом як керованих, так і некерованих факторів;

– відмова технологічного обладнання;

– неправильне функціонування датчиків інформації.

Після усунення цих причин розвідувальний блок, що формує множину оцінок K, перевіряє основні статистичні гіпотези за відповідними критеріями, і, якщо підтверджується нормальність розподілу реалізацій образу, статистична стійкість та однорідність вхідних даних, переводить систему в режим нормального функціонування.

Як критерій оцінки функціональної ефективності навчання СППР застосуємо модифікацію критерію Кульбака [4] для двоальтернативного рішення при допущенні рівноймовірних гіпотез $p(\mu_1) = p(\mu_2) = 0,5$:

$$J_{\Sigma,m} = 0,5\log_2\left(\frac{D_1 + D_2}{\alpha + \beta}\right) * \left[(D_1 + D_2) - (\alpha + \beta)\right] = \\ = \log_2\left(\frac{2 - (\alpha + \beta)}{\alpha + \beta}\right) * \left[1 - (\alpha + \beta)\right], \tag{3}$$

де α, β, D_1, D_2 – точні характеристики навчання СК: помилки першого та другого роду, перша та друга достовірності відповідно.

ПРИКЛАД РЕАЛІЗАЦІЇ АЛГОРИТМУ ФКА

Приклад реалізації алгоритму ФКА із оцінкою статистичних стійкості та однорідності навчальних вибірок розглянемо для апріорного алфавіту із трьох класів, які формувалися за результатами хімічного аналізу вмісту азоту, фосфору та калію при виробництві складного мінерального добрива NPK на ВАТ «Сумихімпром» за такими характеристиками:

– якщо їх відсоток знаходиться в інтервалі 14,3 – 15,5 то реалізація належить до найбільш бажаного класу X_1^o;

– якщо – в інтервалі 13,0 – 15,0, тобто то їх вміст нижче норми, то реалізація належить до класу X_2^o;

– якщо – в інтервалі 14,5 – 16,5, тобто їх вміст вище норми, то реалізація належить до класу X_3^o.

Навчальні матриці складались із 40 векторів-реалізацій, координати яких дорівнювали значенням 41-ї ознаки розпізнавання.

У результаті застосування послідовної оптимізації контрольних допусків на ознаки розпізнавання в рамках алгоритму навчання за МФСВ [8] було побудовано оптимальні контейнери класів розпізнавання для апріорного алфавіту з такими параметрами:

– для класу X_1^o, що характеризує оптимальний заданий технологічний режим, з радіусом контейнера $d_1^* = 8$ кодових одиниць і еталонним вектором-реалізацією

X_1=<11110111111011011110101011111000011111001 0>;

– для класу X_2^o з радіусом $d_2^* = 4$ кодових одиниць й еталонним вектором-реалізацією

X_2=<11110111111111011111101011111101011011110 1>;

– для класу X_3^o з радіусом $d_3^* = 7$ кодових одиниць і еталонним вектором-реалізацією

X_3=<11010111011011001110111001111110111111111 0>.

За умов дослідно-промислових випробувань у цеху складних мінеральних добрив було в режимі КФА проаналізовано системою підтримки прийняття рішень (СППР), яка входить до складу АСКТП виробництва складного мінерального добрива NKP, дискретні значення 530 векторів-реалізацій з кроком квантування за часом 2 хвилини. Із них:

– для 152 реалізацій рішення не приймалось, оскільки вони не відповідали умовам однорідності;

– 89 реалізацій було віднесено до класу віднесено до найбільш бажаного класу X_1^o;

– 95 реалізацій було віднесено до класу віднесено до класу X_2^o;

– 101 реалізацій було віднесено до класу віднесено до класу X_3^o;

– для 93 реалізацій СППР було сформовано дві навчальні матриці для нових класів X_4^o і X_5^o та здійснено перенавчання системи з метою побудови оптимального розбиття для розширеного алфавіту класів.

З метою дослідження причин, за якими 152 реалізації не оброблялися алгоритмом ФКА, на рис. 2 показана динаміка зміни в часі значень 38-ї ознаки (вміст води в пульпі у відсотках).

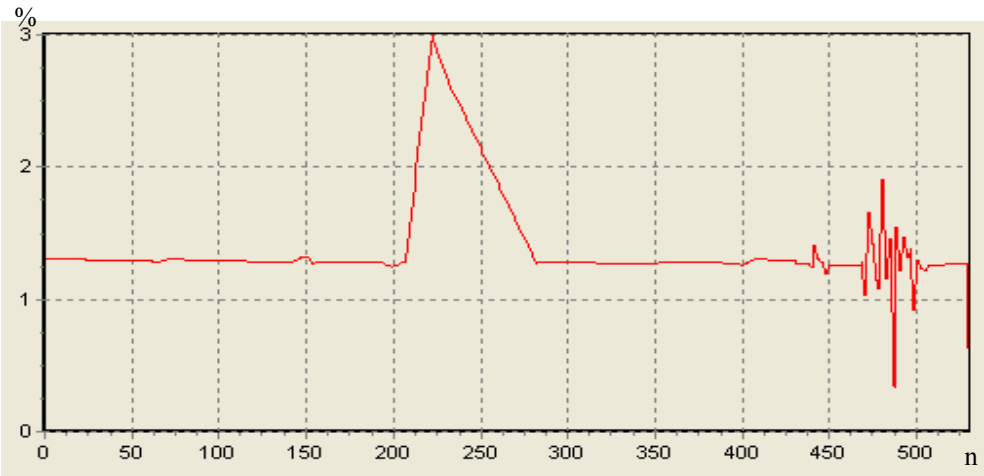


Рисунок 2–Динаміка зміни вмісту води в пульпі за 530 екзаменаційними реалізаціями

Аналіз рис. 2 показує, що на інтервалі часу з 203 по 278 та з 453 по 530 реалізації відбувалися перехідні процеси, що спричинило невідповідність 152 реалізацій умовам нормальності, статистичної стійкості та однорідності.

На рис. 3 і 4 наведено графіки залежності КФЕ (3) від радіуса оптимального гіперсферичного контейнера для нових класів X_4^o і X_5^o відповідно. Тут світлі області графіків позначають робочу область визначення функції КФЕ, в якій здійснюється пошук її глобального

максимуму в процесі оптимізації просторово часових параметрів функціонування СППР.

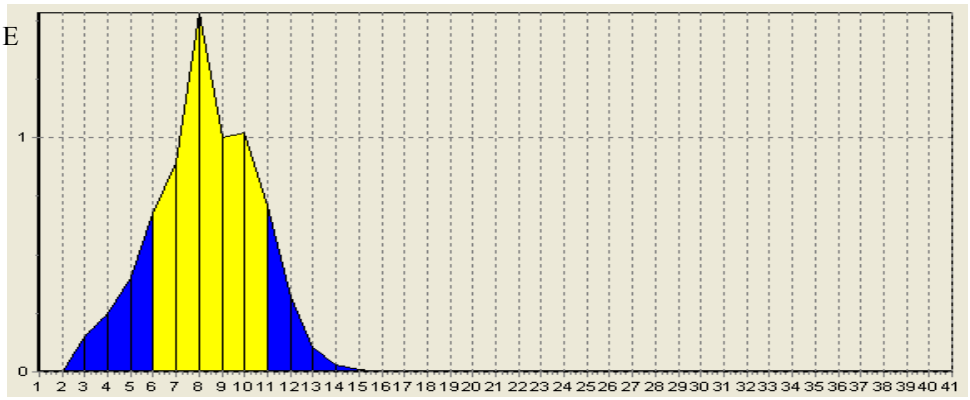


Рисунок 3 – Залежність КФЕ від радіуса контейнера для класу X_4^o

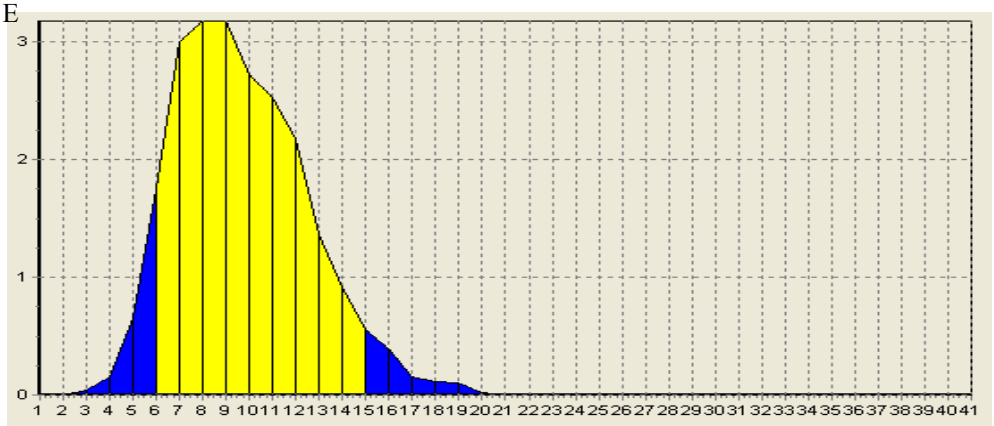


Рисунок 4 – Залежність КФЕ від радіуса контейнера для класу X_5^o

Аналіз рис. 3 і 4 показує, що відновлені в процесі роботи алгоритму ФКА гіперсферичні контейнери нових класів мають такі параметри:

– для класу X_4^o оптимальний радіус $d_4^* = 8$ кодових одиниць (еталонний вектор-реалізація

X_4=<1001010111001011000010011100010010010100000>);

– для класу X_5^o оптимальний радіус $d_5^* = 8$ кодових одиниць (еталонний вектор-реалізація

X_5=<1011100010010100011110111000100100101010000>).

Аналогічно будуються оптимальні контейнери для інших класів, що характеризують допустимі функціональні стани керованого технологічного процесу.

ПЕРСПЕКТИВИ РОЗВИТКУ ТА ЗАСТОСУВАННЯ МЕТОДУ

Подальший розвиток запропонованого в рамках ІЕІТ методу ФКА доцільно здійснювати у таких напрямках:

– розширення ієрархічної структури статистичних критеріїв оцінки навчальних вибірок;

– оптимізація інших параметрів функціонування СППР, таких, як рівні селекції координат двійкових еталонних векторів-реалізацій образу; параметри словника ознак розпізнавання, впливу зовнішнього середовища та інші, з метою побудови безпомилкового за навчальною матрицею класифікатора;

– застосування та подальший розвиток методів завадозахищеного кодування для підтримки асимптотичної достовірності класифікатора за умови розширення алфавіту класів розпізнавання в процесі роботи алгоритму ФКА.

Перспективною областю впровадження одержаних результатів є керовані слабо формалізовані процеси в різних галузях соціально-економічної сфери українського суспільства: нафто-хімічна та металургійна промисловість, нанотехнології, медичне діагностування та інше.

ВИСНОВКИ

1 Запропоновано в рамках ІЕІТ новий метод ФКА, що включає оцінку статистичних характеристик екзаменаційної матриці і дозволяє підвищити ефективність функціонування здатної навчатися СППР шляхом забезпечення однакових умов формування навчальних матриць нових класів розпізнавання.

2 Одержані результати дозволяють адаптувати алгоритми ФКА до реальних технологічних процесів шляхом оцінки статистичних стійкості та однорідності екзаменаційних вибірок й оперативного відключення алгоритму навчання СППР під час перехідних процесів, що відбуваються через вплив як керованих, так і некерованих факторів.

SUMMARY

Propose method factorial classification analysis in the context informational-extreme intelligent technology. Method allowed to organize at the process management faintly formalized process teaching selection new classes, which described by statistical firmness and similarity.

СПИСОК ЛІТЕРАТУРИ

1. Загоруйко Н. Г., Елкина В. Н., Лбов Г. С. Алгоритмы обнаружения эмпирических закономерностей. – Новосибирск: Наука. –1985. – 110 с.
2. Методы анализа данных: Подход, основанный на методе динамических сгущений / Пер. с фр. / Кол. авт. под рук. Э. Дидэ / Под ред. и с предисл. С.А. Айвазяна и В.М. Бухштабера.– М.: Финансы и статистика, 1985.–375 с.
3. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин / Под ред. С.А. Айвазяна.– М.: Финансы и статистика, 1989. – 607 с.
4. Краснопоясовский А.С. Інформаційний синтез інтелектуальних систем керування: Підхід, що ґрунтується на методі функціонально-статистичних випробувань. – Суми: Видавництво СумДУ, 2004.– 261 с.
5. Гублер Е.В. Вычислительные методы анализа и распознавания патологических процес сов. - М.: Медицина, 1978. - 296 с.
6. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. – М.: Наука, 1983. – 416 с.
7. Назаренко О.М. Основи економетрики: Підручник. – Київ: Центр навчальної літератури, 2004. – 392 с.
8. Краснопоясовський А.С., Кий О.М., Волков В.М., Козинець М.В., Шелехов І.В. Класифікаційне управління технологічним процесом виробництва складних мінеральних добрив // Східно-Європейський журнал передових технологій.–2003.– № 6.–С.12–17.

Надійшла до редакції 12 грудня 2006 р.