УДК 681.518:004.93.1′

А.С. ДОВБЫШ, В.В. МОСКАЛЕНКО, А.С. РЫЖОВА

ИНФОРМАЦИОННО-ЭКСТРЕМАЛЬНЫЙ МЕТОД КЛАССИФИКАЦИИ НАБЛЮДЕНИЙ С КАТЕГОРИАЛЬНЫМИ ПРИЗНАКАМИ

Аннотация. Рассмотрен алгоритм информационно-экстремального машинного обучения, основанный на адаптивном кодировании разнотипных первичных признаков распознавания и оптимизации геометрических параметров разбиения пространства вторичных (унифицированных) признаков на классы эквивалентности в процессе итерационного приближения глобального максимума информационного критерия к его граничному значению.

Ключевые слова: распознавание образов, классификатор, машинное обучение, категориальные признаки, система контрольных допусков на признаки распознавания, обучающая матрица, информационный критерий.

ВВЕДЕНИЕ

Во многих практических задачах мониторинга управляемых процессов наблюдения могут быть как с непрерывными (количественными), так и дискретными (называемыми также категориальными или качественными) признаками, принимающими определенные значения из конечного неупорядоченного множества. В таких случаях выявление закономерностей в данных мониторинга затруднено, поскольку подавляющее большинство алгоритмов машинного обучения позволяет учитывать лишь количественные признаки для описания наблюдаемых процессов. Приведение категориальных первичных признаков к количественным вторичным путем простой нумерации значений первичных признаков редко приводит к удовлетворительному результату, поскольку алгоритмы обучения будут учитывать не имеющую смысла упорядоченность. В работах [1–3] предложено использовать Dummy-кодирование категориальных признаков, при котором каждый первичный признак перекодируется в несколько вторичных бинарных признаков конкретно с одной единицей. Подобная перекодировка позволяет применять многие классические алгоритмы обучения, однако это существенно увеличивает размерность пространства признаков, а также накладывает ограничения на структуру признакового описания наблюдений. Другой подход связан с применением в алгоритме машинного обучения метрики перекрытия (расстояния по Хеммингу), при использовании которой степень различия реализаций образов определяется количеством несовпадающих значений первичных признаков [4, 5]. Недостатком таких алгоритмов является игнорирование дополнительной информации, предоставляемой категориальными признаками. Использование метрики перекрытия не позволяет учитывать частоты появления конкретных значений категориальных признаков в конкретном классе распознавания. Одним из решений этой проблемы является применение частотной перекодировки категориальных признаков, что приводит к упорядоченности значений признаков и позволяет применять различные статистические методы [6–8]. При этом по-прежнему остаются актуальными задача построения вычислительно-эффективных решающих правил для случая большого размера входных данных и задача обеспечения высокой обобщающей способности алгоритма машинного обучения для работы с обучающими выборками малого объема. В работах [9–11] эти задачи успешно решаются для данных с количественными признаками путем адаптивного их кодирования двоичным представлением в процессе оптимизации контрольных допусков на значения признаков с применением логарифмических информационных критериев валидации. Применение подобного подхода при наличии категориальных признаков до настоящего времени не рассматривалось.

В данной статье предлагается новый алгоритм информационно-экстремального обучения классификатора с адаптивным кодированием признаков, в котором осуществляется унификация разнотипной информации двоичным представлением, учитывающим вероятностные характеристики как количественных, так и качественных (категориальных) признаков.

ПОСТАНОВКА ЗАДАЧИ

Рассмотрим обучаемый классификатор наблюдений с разнотипными признаками распознавания. Пусть имеем обучающую матрицу $\{y_{m,i}^{(j)},\ i=\overline{1,N};\ j=\overline{1,n};\ m=\overline{1,M}\}$, где N — количество признаков распознавания; n — число наблюдений m-го класса распознавания; M — мощность алфавита классов распознавания $\{X_m^o,\ m=\overline{1,M}\}$, характеризующих функциональное состояния контролируемого процесса. Кроме того, известна структура вектора параметров обучения классификатора

$$g=\langle \delta_i, d_m\rangle, \tag{1}$$

где δ_i — параметр поля контрольных допусков для i-го признака; d_m — радиус вписанного в гиперпараллелепипед с единичными ребрами гиперсферического контейнера класса X_m^o, восстанавливаемого в радиальном базисе бинарного пространства признаков $\Omega_б$. При этом заданы ограничения на параметры обучения: $1\le d_m < d(x_m \oplus x_c)$; здесь $d(x_m \oplus x_c)$ — кодовое расстояние от эталонного вектора x_m класса X_m^o к эталонному вектору x_c ближайшего (соседнего) класса $X_c^o \in \{X_m^o\}$; $0\le \delta_i \le \delta_{\max,i}$, где $\delta_{\max,i}$ — максимально допустимое значение параметра поля контрольных допусков для значения i-го признака распознавания.

Необходимо в процессе обучения классификатора найти оптимальные значения координат вектора параметров (1), обеспечивающих максимальное значение усредненного по алфавиту классов распознавания информационного критерия функциональной эффективности (КФЭ):

$$\overline{E}^* = \frac{1}{M}\sum_{m=1}^{M}\max_{\{k\}} E_m^{(k)}, \tag{2}$$

где $E_m^{(k)}$ — вычисляемый на k-й итерации информационный КФЭ обучения классификатора для распознавания реализации класса X_m^o; $\{k\}$ — множество итераций максимизации КФЭ (множество шагов обучения).

При функционировании классификатора в рабочем режиме (экзамена) необходимо определить принадлежность распознаваемой реализации одному из классов $\{X_m^o\}$, сформированного на этапе обучения алфавита.

АЛГОРИТМ ОБУЧЕНИЯ

Процедура кодирования признаков в рамках информационно-экстремального обучения начинается с вычисления верхнего и нижнего контрольных допусков при заданном параметре δ_i, $i=\overline{1,N}$ (рис. 1). При этом предварительно для каждого количественного признака вычисляется его выборочное среднее значение $\overline{y}_{m,i}$, а для категориального признака — относительная частота появления каждого его значения $\{f_{m,i,\upsilon},\ \upsilon=\overline{1,V_i}\}$, где V_i — количество дискретных значений признака.

На рис. 1 показаны верхний $A_{в,i}$ и нижний $A_{н,i}$ контрольные допуски для i-го количественного признака, которые определяются относительно среднего выборочного значения в базовом классе $X_б^o \in \{X_m^o\}$:

$$A_{в,i}=\overline{y}_{б,i}+\delta_i;\ \ A_{н,i}=\overline{y}_{б,i}-\delta_i. \tag{3}$$

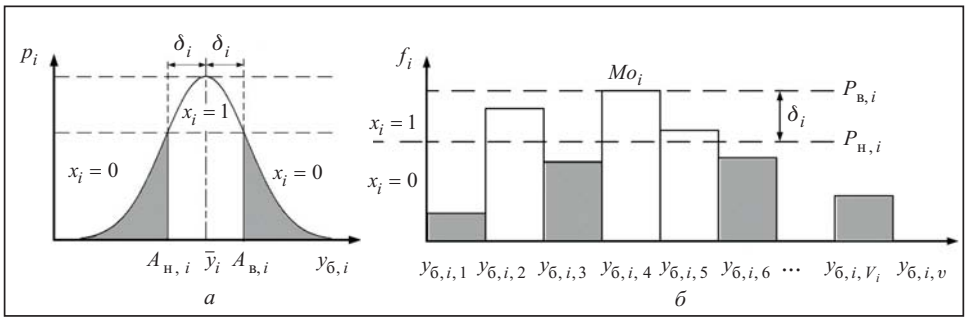

Рис. 1. График изображения контрольных допусков на значение *i*-го количественного (*а*) и категориального (*б*) признаков базового класса

Для каждого категориального признака верхний контрольный допуск $P_{в,i}$ равен максимальной (модальной) относительной частоте появления значения признака в базовом классе $X_б^o \in \{X_m^o\}$, а нижний контрольный допуск $P_{н,i}$ определяется значением параметра поля допуска:

$$P_{в,i} = f_{б,i,Mo};\ P_{н,i} = f_{б,i,Mo} - \delta_i, \tag{4}$$

где $f_{б,i,Mo} = \max\limits_{v} \{f_{m,i,v}\}$ — максимальная относительная частота появлений значения (частота моды Mo_i) *i*-го категориального признака в базовом классе.

Соответственно формирование бинарной обучающей матрицы $\{x_{m,i}^{(j)}, i = \overline{1, N};\ j = \overline{1, n};\ m = \overline{1, M}\}$ осуществляется кодированием количественных и категориальных первичных признаков по правилам

$$x_{m,i}^{(j)} = \begin{cases} 1, & \text{если } A_{н,i} \le y_{m,i}^{(j)} \le A_{в,i}, \\ 0 & \text{в противном случае;} \end{cases} \tag{5}$$

$$x_{m,i}^{(j)} = \begin{cases} 1, & \text{если } P_{н,i} \le f_{m,i}^{(j)} \le P_{в,i}, \\ 0 & \text{в противном случае,} \end{cases} \tag{6}$$

где $y_{m,i}^{(j)}$ — значение *i*-го количественного признака в *j*-й реализации класса X_m^o; $f_{m,i}^{(j)}$ — относительная частота появлений значения $y_{m,i}^{(j)}$ *i*-го категориального признака в реализациях класса X_m^o.

Квазиоптимизация параметра $\delta = \delta_i$, $i = \overline{1, N}$, контрольного допуска предназначена для определения стартовых значений признаков распознавания, используемых в алгоритме их последовательной оптимизации, и осуществляется по итерационной процедуре [9]

$$\delta^* = \arg\max_{G_\delta} \left\{ \frac{1}{M} \sum_{s=1}^{M} \left[\max_{G_E \cap G_d} E_m \right] \right\}, \tag{7}$$

где G_δ — область допустимых значений параметра контрольного допуска; G_d — область допустимых значений радиуса гиперсферического контейнера; G_E — рабочая область определения функции КФЭ.

Последовательная оптимизация параметра δ_i системы контрольных допусков для *i*-го признака осуществляется по итерационной процедуре [10]

$$\delta_i^* = \arg \left\{ \bigotimes_{l=1}^{L} \max_{G_{\delta_i}} \left\{ \frac{1}{M} \sum_{m=1}^{M} \left[\max_{G_E \cap G_d} E_m^{(l)} \right] \right\} \right\}, \tag{8}$$

где $E_m^{(l)}$ — КФЭ обучения классификатора для распознавания реализации m-го класса на l-м шаге последовательной процедуры оптимизации; G_{δ_i} — область допустимых значений параметра контрольных допусков для i-го признака; $\otimes$ — символ операции повтора; L — количество прогонов итерационной процедуры последовательной оптимизации параметра контрольных допусков.

В качестве КФЭ обучения классификатора рассмотрим модифицированную информационную меру Кульбака [11], в которой отношение правдоподобия представлено в виде отношения полной вероятности правильного принятия решений P_{true} к полной вероятности ошибочного принятия решений P_{false}. Для случая равновероятных и двухальтернативных гипотез мера Кульбака имеет вид

$$E_m = [P_{\text{true},m} - P_{\text{false},m}]\log_2 \frac{P_{\text{true},m}}{P^{(k)}_{\text{false},m}} = \left| \begin{array}{l} P_{\text{true},m} = 0.5D_{1,m} + 0.5D_{2,m}; \\ P_{\text{false},m} = 0.5\alpha_m + 0.5\beta_m; \\ \alpha_m = 1 - D_{1,m};\ D_{2,m} = 1 - \beta_m; \\ D_{1,m} = \dfrac{K_{1,m}}{n};\ \beta_m = \dfrac{K_{2,m}}{n} \end{array} \right| =$$

$$= \frac{1}{n}[K_{1,m} - K_{2,m}] \cdot \log_2 \left(\frac{n + (K_{1,m} - K_{2,m}) + 10^{-r}}{n - (K_{1,m} - K_{2,m}) + 10^{-r}} \right), \qquad (9)$$

где $D_{1,m}$ — первая достоверность для m-го класса; $D_{2,m}$ — вторая достоверность; α_m — ошибка первого рода; β_m — ошибка второго рода; n — количество реализаций в обучающей выборке класса X_m^o; $K_{1,m}$ — количество событий, характеризующих попадание реализаций в контейнер класса X_m^o, если они действительно являются реализациями этого класса; $K_{2,m}$ — количество событий, характеризующих попадание реализаций в контейнер класса X_m^o, если они в действительности принадлежат другому классу.

Нормированная модификация критерия (9) представлена в виде

$$\hat{E}_m = \frac{E_m}{E_{\max}}, \qquad (10)$$

где $E_{\max}$ — максимальное значение критерия, полученное при $K_{1,m} = n$ и $K_{2,m} = 0$.

В качестве рабочей (допустимой) принимается область определения функции информационного КФЭ, в которой первая и вторая достоверности должны соответственно удовлетворять неравенствам $0.5 \le D_1 \le 1$ и $0.5 \le D_2 \le 1$.

Определение принадлежности тестовой реализации $x^{(t)}$ к классу X_m^o основано на анализе значений функции принадлежности

$$\mu_m = 1 - \frac{d(x_m^* \oplus x^{(t)})}{d_m^*}, \qquad (11)$$

где $d(x_m^* \oplus x^{(t)})$ — кодовое расстояние между эталонным вектором x_m^* и распознаваемой реализацией $x^{(t)}$; d_m^* — оптимальный радиус контейнера класса X_m^o.

Для реализации процедуры экзамена необходимо сформировать двоичное представление $x^{(t)}$ для распознаваемой реализации $y^{(t)}$. Для этого в памяти хранятся значения границ контрольных допусков для i-го количественного признака

$A_{\text{н},i}$ и $A_{\text{в},i}$, а для категориального признака — множество значений первичного признака, кодирующихся как единица при оптимальных контрольных допусках $P_{\text{н},i}$ и $P_{\text{в},i}$.

Таким образом, суть информационно-экстремального машинного обучения состоит в итерационном приближении глобального максимума информационного КФЭ к его граничному значению в процессе адаптивного кодирования разнотипных первичных признаков распознавания путем оптимизации геометрических параметров разбиения пространства вторичных (унифицированных) признаков на классы эквивалентности.

РЕЗУЛЬТАТЫ ФИЗИЧЕСКОГО МОДЕЛИРОВАНИЯ

Рассмотрим результаты реализации предложенного алгоритма на примере обучения системы поддержки принятия решений в составе автоматизированной системы управления выращиванием крупногабаритных сцинтилляционных монокристаллов с расплава [12]. По результатам анализа архивной базы записей о кратковременных отказах и других флагах изменения технического состояния функциональных модулей (контроллеров) системы управления ростовым процессом сформирована многомерная обучающая матрица [12, 13]. Отметим, что векторы-реализации классов состоят из десяти категориальных признаков, описывающих техническое состояние контроллеров, и шести количественных признаков, получаемых от датчиков физических величин. Алфавит классов распознавания под воздействием технического состояния контроллеров и внешних факторов квалифицировал условия изменения выпуклости фронта кристаллизации по системе оценок: меньше нормы, норма и больше нормы.

На рис. 2 показано распределение значений каждого из десяти категориальных признаков по трем классам распознавания. У гистограммы количество столбцов s_i, $i=\overline{1, N}$, соответствует количеству V_i дискретных значений i-го категориального признака. Сетчатая штриховка соответствует классу X_1^o — норма, кружевная штриховка соответствует классу X_2^o — больше нормы, незаштрихованная часть соответствует классу X_3^o — меньше нормы.

Рассмотрим процесс оптимизации параметров функционирования информационно-экстремального классификатора. На рис. 3 показан график итерационного процесса оптимизации параметра контрольных допусков согласно процедуре (8), а на рис. 4 — график значений контрольных допусков для частот встречаемости значений категориальных признаков.

Из рис. 3 следует, что в процессе 1070 итераций (количество изменений системы контрольных допусков) алгоритма оптимизации получена оптимальная система контрольных допусков с вектором параметров $\{\delta_i^*,\ i=\overline{1, N}\}$, обеспечивающая граничное значение глобального максимума усредненного нормированного КФЭ (см. (10)) $\overline{E}^*=1.00$. Исходя из этого можно утверждать, что в процессе обучения построены безошибочные по обучающей матрице решающие правила, гарантирующие принятие решений в режиме экзамена с высокой достоверностью.

На рис. 5 показан процесс оптимизации геометрических параметров разбиения пространства вторичных (унифицированных) признаков на классы эквивалентности при оптимальной системе контрольных допусков. Как видим, оптимальные значения радиусов гиперсферических контейнеров классов X_1^o, X_2^o и X_3^o соответственно равны в кодовых единицах $d_1^*=5$, $d_2^*=3$, $d_3^*=2$.

Для оценки функциональной эффективности разработанного информационно-экстремального классификатора в режиме экзамена проводилось сравнение результатов кросс-валидации его решающих правил по сформированной тестовой

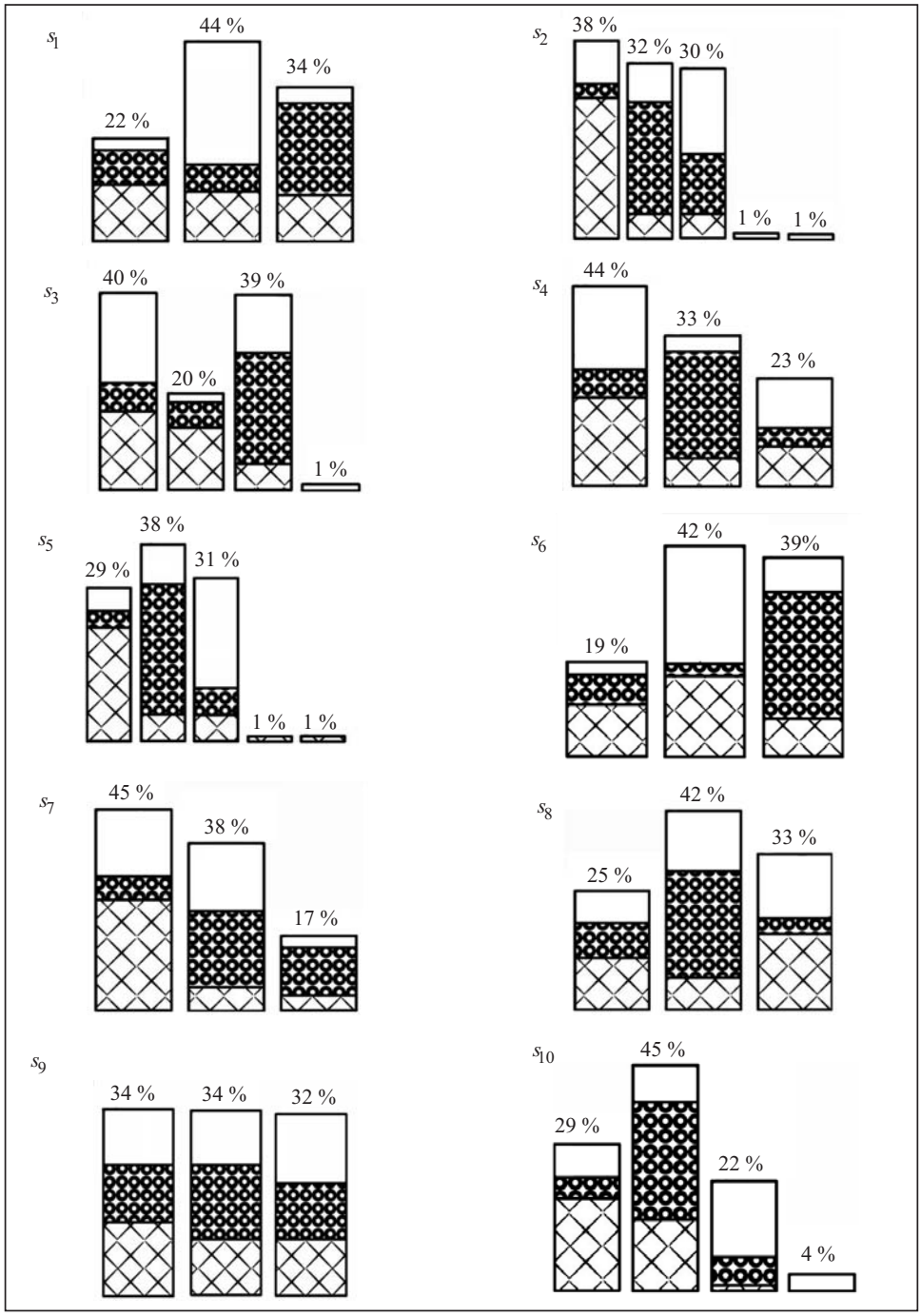

Рис. 2. График распределения значений категориальных признаков по трем классам распознавания

обучающей матрице с результатами, полученными при использовании пакета Weka [14]. В табл. 1 приведены значения полных усредненных по алфавиту классов распознавания вероятностей правильного принятия решений $\overline{P}_{\text{true}}$, неправильного принятия решений $\overline{P}_{\text{false}}$ и времени обучения. Точность классификаторов, входящих в пакет Weka, у которых для определения степени различия наблюдений с категориальными признаками используется метрика перекрытия (расстоя-

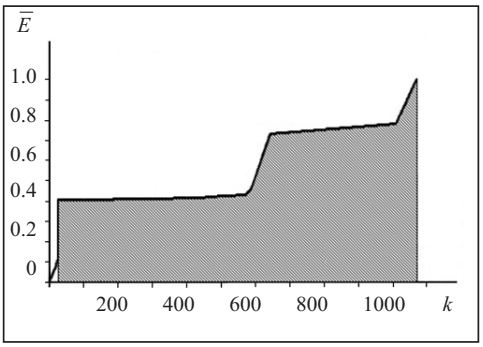

Рис. 3. График зависимости усредненного значения КФЭ (10) от количества итераций оптимизации системы контрольных допусков

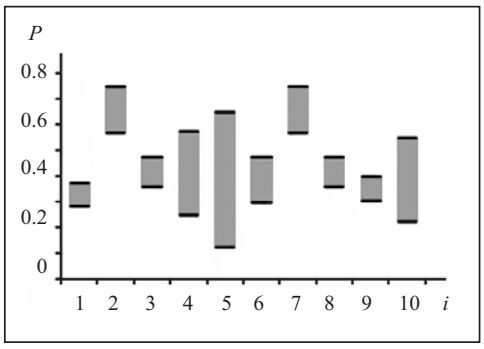

Рис. 4. График оптимальных верхнего и нижнего контрольных допусков для частот встречаемости значений категориальных признаков

ние по Хеммингу), уступает информационно-экстремальному классификатору, который при этом не уступает рассмотренным классификаторам по оперативности процесса обучения.

Таким образом, частотное перекодирование категориальных признаков позволяет использовать информационно-экстремальное обучение для построения четкого разбиения пространства вторичных (унифицированных) признаков на классы эквивалентности путем оптимизации в информационном смысле контрольных допусков на значения количественных признаков распознавания и контрольных допусков на значения частот появления значений категориальных признаков распознавания.

ЗАКЛЮЧЕНИЕ

Рассмотренный в статье алгоритм информационно-экстремального машинного обучения позволяет получить вычислительно эффективные решающие правила с высокой обобщающей способностью по обучающей матрице с разнотипными признаками.

Физическое моделирование по данным архивной базы записей о флагах технического состояния функциональных модулей и показаний датчиков системы управления выращиванием монокристаллов показало преимущество разработанного алгоритма машинного обучения по сравнению с другими известными алгоритмами. При этом удалось получить безошибочные по обучающей матрице реша-

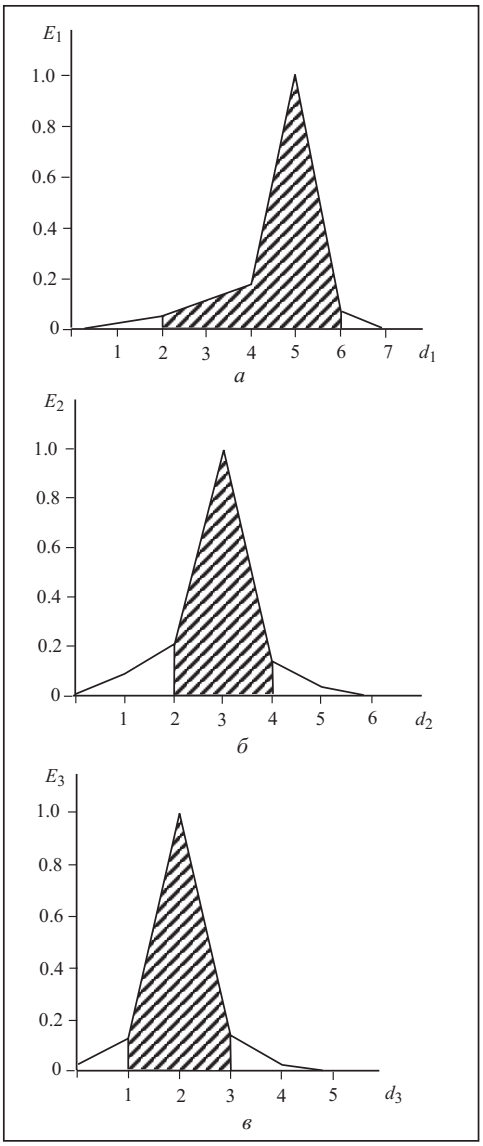

Рис. 5. Графики зависимости КФЭ E_n от радиуса d_n контейнера класса X_1^o (*а*); класса X_2^o (*б*); класса X_3^o (*в*)

Таблица 1

Название классификатора	Значение полной вероятности		Время обучения, с
	$\overline{P}_{\text{true}}$	$\overline{P}_{\text{false}}$	
Наивный Байес	0.849	0.076	0.07
Машина опорных векторов	0.966	0.017	0.17
Деревья решений J48	0.916	0.042	0.21
Многослойный персептрон	0.992	0.004	0.86
Сеть радиально-базисных функций	0.874	0.063	0.13
Информационно-экстремальный	1.000	0.000	0.09

ющие правила, что позволяет повысить точность распознавания условий формирования ошибочного управления и обеспечить своевременность коррекции технологического процесса.

СПИСОК ЛИТЕРАТУРЫ

1. Cerioli A., Riani M., Atkinson A.C. Robust classification with categorical variables // Proceedings in Computational Statistics. — Heidelberg: Physica-Verlag HD, 2006. — P. 507–519.
2. Alkharusi H. Categorical variables in regression analysis: A comparison of dummy and effect coding // International Journal of Education. — 2012. — **4**, N 2. — P. 202–210.
3. Shyu M.-L., Kuruppu-Appuhamilage I., Chen S.-C., Chang L.W., Goldring T. Handling nominal features in anomaly intrusion detection problems // 15th Workshop on Research Issue in Data Engineering: Stream Data Mining and Applications. — Harvard: IEEE Computer Society, 2005. — P. 55–62.
4. Дьяконов А.Г. Методы решения задач классификации с категориальными признаками // Прикладная математика и информатика. — 2014. — № 46. — С. 81.
5. Wilson D.R., Martinez T.R. Improved heterogeneus distance functions // Journal of Artificial Intelligence Research. — 1997. — **6.** — P. 1–34.
6. Ienco D., Pensa R.G., Meo R. Context-based distance learning for categorical data clustering // Advances in Intelligent Data Analysis VIII. Lecture Notes in Computer Science. — 2009. — **5772**. — P. 83–94.
7. Янковская А.Е., Берестнева О.Г., Муратова Е.А. Извлечение знаний с применением алгоритма адаптивного кодирования разнотипной информации // Штучний інтелект. — 2002. — № 2. — С. 315–322.
8. Кандрашова Н.В., Павлов В.А., Павлов А.В. Метод Байеса и МГУА в задаче классификации с переменными разного типа // Матеріали міжнародної науково-технічної конференції «Геоінформаційні системи і комп'ютерні технології еколого-економічного моніторингу». — Дніпропетровськ: ГВУЗ НГУ, 2014. — С. 4.
9. Dovbysh A.S., Budnyk N.N., Moskalenko V.V. Information-extreme algorithm for optimizing parameters of hyperellipsoidal containers of recognition classes // Journal of Automation and Information Sciences. — 2012. — **44**, N 10. — P. 35–44.
10. Довбиш А.С. Основи проектування інтелектуальних систем. — Суми: СумДУ, 2009. — 171 с.
11. Москаленко В.В., Довбиш А.С. , Рижова А.С. Інтелектуальна автоматизована система керування з оптимізацією часових параметрів аналізу вхідних даних // Вісник Сумського державного університету. — 2013. — № 3. — С. 7–14.
12. Суздаль В.С., Стадник П.Е., Герасимчук Л.И., Епифанов Ю.М. Сцинтилляционные монокристаллы: автоматизированное выращивание. — Харьков: ИСМА, 2009. — 260 с.
13. Sipos R., Fradkin D., Moerchen F., Wang Z. Log-based predicitive maintenance // Proceeding of 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — New York: ACM, 2014. — P. 1867–1876.
14. Weka. waikato.ac.nz.

Поступила 30.01.2015