УДК 681.518:004.93.1’

**В.В. Москаленко**

Сумской государственный университет, г. Сумы, Украина

# ИНФОРМАЦИОННО-ЭКСТРЕМАЛЬНОЕ МАШИННОЕ ОБУЧЕНИЕ ПО НЕСБАЛАНСИРОВАННЫМ ДАННЫМ БОЛЬШОГО РАЗМЕРА

Предлагается алгоритм информационно-экстремального машинного обучения с итерационным сбором обучающей выборки, позволяющий производить обучение классификатора на исходных данных большого обьема без их дополнительной предобработки (нормализации, балансировки, фильтрации). В качестве примера рассмотрено реализацию предложенного алгоритма для обучения классификатора по большим несбалансированным наборам архивных данных мониторинга сетевого трафика и технологического процесса выращивания сцинтиляционных монокристаллов.

НЕСБАЛАНСИРОВАННЫЕ ВЫБОРКИ, ИНФОРМАЦИОННЫЙ КРИТЕРИЙ, РЕЦЕПТИВНЫЕ ПОЛЯ ПРИЗНАКОВ, МАШИННОЕ ОБУЧЕНИЕ, КЛАССИФИКАТОР

## Введение

В практических условиях мониторинга управляемых процессов часто накапливаются большие объемы архивной истории со значительной несбалансированностью набора реализаций классов распознавания, что усложняет применение традиционных методов машинного обучения [1, 2]. Одной из простых техник учета несбалансированности является увеличение “премий” за правильную и увеличение “штрафов” за ошибочную классификации реализаций миноритарных классов [1]. Это увеличивает чувствительность к редким событиям (например: редкие болезни, сетевые атаки, аномальные состояния технологического процесса и т.п.), распознавание которых имеет большую ценность, однако одновременно уменьшается суммарная точность распознавания. В связи с этим в последнее время большую популярность приобрели методы сэмплирования under-sampling и oversampling [3]. Применение undersampling позволяет сбалансировать число реализаций различных классов случайной выборкой из мажоритарного класса, однако это может привести к потере важной информации. Методы oversampling искусственно увеличивают количество реализаций миноритарного класса путем дублирования или путем их генерирования с помощью малых отклонений от реальных реализаций миноритарного класса. При этом в обоих случаях могут быть искажены вероятностное распределение и другие закономерности. Для повышение точности классификации реализаций несбалансированных классов распознавания также применяют итерационный сбор сбалансированной обучающей выборки (bootstrap) или методы последовательной (boosting) либо параллельной (bagging) композиции ансамбля классификаторов, обученных на различных подвыборках. Эти подходы подразумевают многократное формирование подвыборок и вызова процедуры обучения и могут быть реализованы в распределенной вычислительной среде. При этом в случае ансамбля классификаторов получается большая модель комбинированого классификатора, а в случае итерационного сбора сложность модели определяется используемым алгоритмом машинного обучения.

В условиях “экстремальной” несбалансированности обучающих данных (объем мажоритарного класса превышает обьем миноритарного больше чем в сотни раз) и ограниченного времени анализа большого обьема входных данных предпочтительнее использовать вычислительно эффективные классификаторы с большой обобщающей способностью. Как было показано в работах [4, 5], наибольшим быстродействием обладают классификаторы с грубым кодированием признаков на основе рецептивных полей, поскольку позволяют трансформировать пространство признаков с помощью простых операций сравнения. При этом гарантированную точность классификации при малых размерах обучающей выборки миноритарного класса могут обеспечить информационно-екстремальные алгоритмы благодаря “сглаживающему” эффекту логарифмических информационных мер [5, 6].

В статье предложено алгоритм информационно-екстремального машиного обучения и модификацию алгоритма итерационного сбора обучающей выборки, что позволяет обучать классификатор по исходных данных большого обьема без их дополнительной предобработки (нормализации, балансировки, фильтрации) и получать простые рещающие правила, использующие минимум памяти и вычислительных ресурсов.

## 1. Постановка задачи

Рассмотрим систему поддержки принятия решений, в состав которой входит обучающийся классификатор. Пусть архивная база данных имеет большой обьем классифицированных наблюдений за управляемым процессом $\{u_{m,i}^{(t)} \mid i=\overline{1,N}; t=\overline{1,T_m}; m=\overline{1,M}\}$, где $N$ – количество признаков распознавания; $T_m$ – количество наблюдений функционального состояния,

соответствующего $m$-му классу распознавания; $M$ – мощность алфавита классов распознавания $\{X_m^o \mid m=\overline{1,M}\}$. Кроме этого, известен структурированный вектор параметров обучения классификатора $g=<\delta_i, d_m>$, где $\delta_i$ – параметр, определяющий полуширину симметрического рецептивного поля для $i$-го признака распознавания относительно среднего значения $i$-го признака в базовом классе распознавания $X_1^o$; $d_m$ – радиус гиперсферического (вписанного в единичный гиперпараллелепипед) контейнера класса $X_m^o$, восстанавливаемого в процессе обучения в радиальном базисе бинарного пространства признаков $\Omega_Б$. При этом заданы ограничения на параметры обучения: $d_m \in [1; d(x_m \oplus x_c)-1]$, где $d(x_m \oplus x_c)$ – кодовое расстояние от эталонного вектора $x_m$ класса $X_m^o$ к эталонному вектору $x_c$ ближайшего (соседнего) класса $X_c^o \in \{X_m^o\}$; $\delta_i \in [0;0{,}5 \cdot \delta_{H,i}]$, где $\delta_{H,i}$ – максимально допустимая ширина рецептивного поля для $i$-го признака распознавания.

Необходимо сформировать репрезентативную обучающую выборку $\{y_{m,i}^{(j)} \mid j=1,n_m^*\} \subset \{u_{m,i}^{(k)}\}$ и в процессе (до)обучения классификатора найти оптимальные значения координат вектора параметров $g^*$, обеспечивающих максимальное значение усреднённого по алфавиту классов распознавания информационного критерия функциональной эффективности (КФЭ)

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

где $E_m$ – информационный КФЭ обучения классификатора распознавать реализации класса $X_m^o$; $\{k\}$ – упорядоченное множество шагов обучения.

При функционировании классификатора в режиме скользящего контроля необходимо найти неверно распознанные реализации на полном наборе априорно классифицированных данных $\{u_{m,i}^{(k)}\}$ для формирования дообучающего набора данных.

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

### 2. Алгоритм обучения классификатора

Процедура грубого кодирования признаков в рамках информационно-экстремального обучения начинается с вычисления левой и правой границы рецептивных полей соответственно:

$$A_{Л,i} = \overline{y}_{1,i} - \delta_i;\ A_{П,i} = y_{1,i} + \delta_i,$$

где $\overline{y}_{1,i}$ – выборочное среднее значение $i$-го признака распознавания по реализациях базового класса $X_1^o$.

Квазиоптимизация параметра $\delta = \delta_i, i=\overline{1,N}$ рецептивного поля предназначена для определения стартовых значений, которые соответствуют рабочей области определения функции информационного КФЭ, и осуществляется по итерационной процедуре

$$\delta^* = \arg\max_{G_\delta}\left\{\frac{1}{M}\sum_{s=1}^{M}\left[\max_{G_E \cap G_d} E_m\right]\right\}, \qquad (2)$$

где $G_\delta$ – область допустимых значений параметра рецептивного поля; $G_d$ – область допустимых значений радиуса гиперсферического контейнера; $G_E$ – рабочая область определения функции КФЭ.

Последовательная оптимизация параметра $\delta_i$ рецептивного поля для $i$-го признака осуществляется по итерационной процедуре

$$\delta_i^* = \arg\left\{\bigotimes_{l=1}^{L}\max_{G_{\delta_i}}\left\{\frac{1}{M}\sum_{m=1}^{M}\left[\max_{G_E \cap G_d} E_m^{(l)}\right]\right\}\right\}, \qquad (3)$$

где $E_m^{(l)}$ – КФЭ обучения классификатора распознавать реализации $m$-го класса на $l$-м шаге последовательной процедуры оптимизации; $G_{\delta_i}$ – область допустимых значений параметра рецептивного поля для $i$-го признака; $\otimes$ – символ операции повтора; $L$ – количество прогонов итерационной процедуры последовательной оптимизации параметра рецептивных полей.

Базовый алгоритм [5,6] является вложенным в процедуры (2) и (3) и осуществляет построение контейнеров для каждого класса распознавания.

В качестве КФЭ обучения классификатора рассмотрим модификацию информационной меры Кульбака [6], в которой отношение правдоподобия представлено в виде отношения полной веротности правильного принятия решений $P_{true}$ к полной вероятности ошибочного принятия решений $P_{false}$. Для случая несбалансированных выборок и двухальтернативных гипотез мера Кульбака имеет видм

$$E_m^{(k)} = \left[P_{true,m}^{(k)} - P_{false,m}^{(k)}\right]\log_2\frac{P_{true,m}^{(k)}}{P_{false,m}^{(k)}} =$$

$$= \left|\begin{array}{l} P_{true,m}^{(k)} = p_1 D_{1,m} + p_2 D_{2,m} \\ P_{false,s}^{(k)} = p_1\alpha_m + p_2\beta_m \\ p_1 = \dfrac{n_m}{n_m+n_c}; p_2 = \dfrac{n_m}{n_m+n_c} \\ \alpha_m = 1 - D_{1,m}; D_{2,m} = 1-\beta_m \\ \tilde{D}_{1,m} = \dfrac{K_{1,m}}{n_m}; \tilde{\beta}_m = \dfrac{K_{2,m}}{n_c} \end{array}\right| =$$

$$= \frac{\left[n_c - n_m + 2\cdot(K_{1,m}^{(k)} - K_{2,m}^{(k)})\right]}{n_m+n_c} * \log_2\left(\frac{n_c + (K_{1,m}^{(k)} - K_{2,m}^{(k)}) + 10^{-r}}{n_m - (K_{1,m}^{(k)} - K_{2,m}^{(k)}) + 10^{-r}}\right), \qquad (4)$$

где $D_{1,m}^{(k)}$ – первая достоверность, вычисленная на $k$-м шагу обучения для $m$-го класса; $D_{2,m}^{(k)}$ – вторая достоверность; $\alpha_m^{(k)}$ – ошибка второго рода; $\beta_m{}^{(k)}$ – ошибка второго рода; $n_m$ – количество реализаций

в обучающей выборке класса $X_m^o$; $n_c$ – количество реализаций, пренадлежащих "соседнему" классу; $K_{1,m}$ – количество событий, харатеризирующих принадлежность реализаций к контейнеру класса $X_m^o$, если они действительно являются реализациями этого класса на $k$ -м шаге обучения; $K_{2,m}$ – количество событий, характеризирующих принадлежность реализаций контейнеру класса $X_m^o$, если они на самом деле принадлежат другому классу.

При расчете точностных характеристик контейнера класса $X_m^o$ в качестве соседнего выбирают класс $X_c^o$ с эталонным вектором $x_c = \arg\min\{d(x_m \oplus x_c)\}$ при $m \neq c$ или согласно принципу $k$-ближайших "соседей" соседним образом можно считать набор $n_m$ ближайших реализаций

$$\{x_c^{(j)} \mid j = \overline{1, n_m}\} \in \left[\bigcup_{c=1}^{M} X_c^o\right] \setminus X_m^o .$$

Нормированная модификация критерия (4) представлена в виде

$$\hat{E}_m^{(k)} = \frac{E_m^{(k)}}{E_{\max}} , \qquad (5)$$

где $E_{\max}$ – максимальное значение критерия, полученное при $K_{1,m}^{(k)} = n_m$ и $K_{2,m}^{(k)} = 0$.

При этом рабочая (допустимая) область определения функции информационного КФЭ ограниченна неравенствами $D_1 \geq 0{,}5$ и $D_2 \geq 0{,}5$.

Определение принадлежности тестовой реализации $x^{(t)}$ к классу $X_m^o$ основано на анализе значений функции принадлежности, которая имеет такой простой вид

$$\mu_m = 1 - \frac{d(x_m^* \oplus x^{(t)})}{d_m^*} , \qquad (6)$$

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 $X_m^o$.

Основная идея обучения при большом и несбалансированном объеме данных в рамках информационно-экстремальной технологии состоит в начальном обучении на малых случайных выборках реализаций классов (одинакового размера) и в постепенном улучшении полученных решающих правил в процессе скользящего контроля на полном наборе данных. Неверно распознанные реализации классов используются для последовательного дополнения обучающей выборки. При этом стартовыми значениями параметров рецептивных полей при дообучении за последовательной процедурой оптимизации (3) являются оптимальные значения параметров, полученные на предыдущей итерации изменения обучающей выборки. Процесс поиска неверно распознанных реализаций классов можно распараллелить (рис. 1) с использованием технологий распределенной обработки, например при терабайтных объемах данных (Big Data) наиболее эффективной является инфраструктура Hadoop [7].

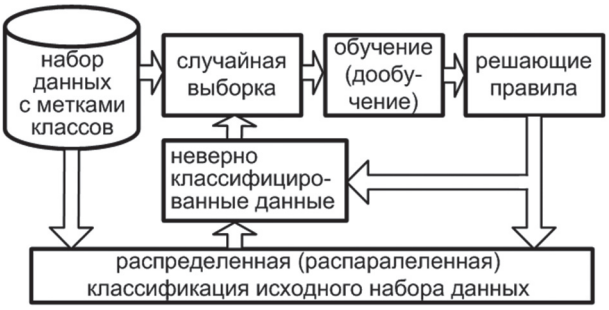



Рис. 1. Схема обучения на большом объеме данных

В отличии от оригинального bootstrap-алгоритма в даном подходе пропуск полудублирующих (почти совпадающих) реализаций осуществляется путем предварительного тестирования решающих правил на каждой добавляемой реализации с последующим дообучением в случае неверной классификации [1]. Это позволяет увеличить разнообразие реализаций образов не допуская "экстремальной" несбалансированости. При этом высокая обобщающая способность информационно-экстремальных решающих правил должна обеспечить сокращение количества итераций дообучения, а вычислительно эффективные решающие правила обеспечивают быстрое сканирование (при скользящем контроле) набора данных.

Основные преимущества предложенного подхода можно исследовать на примере обучения по обучающим данным с обьемом на порядок меньшим по сравнению с Big Data. Для этого рассмотрим следующую модификацию алгоритма обучения:

Шаг 1. Инициализация массивов оптимального $\delta_i^*$ и стартового $\delta_i^{start}$ параметров рецептивных полей: $\delta_i^* := 0$; $\delta_i^{start} := 0$ при $i = \overline{1, N}$.

Шаг 2. Формирование случайной выборки $Y = \{y_{m,i}^{(j)} \mid m = \overline{1, M}; j = \overline{1, n_{\min}}; i = \overline{1, N}\}$, где $n_{\min}$ – минимальный по умолчанию обьем выборки, из априрно классифицированного набора данных $Y_{test} = \{u_i^{(t)} \mid t = \overline{1, T};\ i = \overline{1, N}\}$, использующегося в дальнейшем для проверки качества решающих правил.

Шаг 3. Запуск информационно-эктремального обучения по обучающей матрице $Y$ с квазиоптимизацией параметра рецептивных полей $\delta = \delta_i$ за процедурой (2) при стартовых параметрах $\{\delta_i^{start}\}$.

Шаг 4. $\delta_i^{start} := \delta^*$.

Шаг 5. Инициализация счетчика векторов-реали-заций тестового набора данных: $t := 0$.

Шаг 6. Запуск информационно-эктремального обучения с последовательной оптимизацией параметра рецептивных полей $\{\delta_i\}$ по процедуре (3) при стартовых параметрах $\{\delta_i^{start}\}$.

Шаг 7. $\delta_{s_i}^{start} := \delta_{s_i}^*$.

Шаг 8. $t := t+1$.

Шаг 9. Если $t \le T$ определить принадлежность реализации $y^{(t)}$ к одному из классов алфавита $\{X_m^o\}$ по максимальному значению функции (6), иначе переход к шагу 11.

Шаг 10. Если принадлежность вектора-реализации $y^{(t)}$ не совпадает с меткой априорной классификации, то добавить $y^{(t)}$ к обучающей матрице $Y$ и перейти к шагу 6, иначе к шагу 8.

Шаг 11. ОСТАНОВ.

Таким образом, алгоритм информационно-экстре-мального машинного обучения по большим несбалансированным данным состоит в итерационном приближении глобального максимума информационного КФЭ к его граниченому значению в процессе итерационного сбора обучающих реализаций и оптимизации параметров обучения. При этом итерационное формирование обучающей матрицы позволяет проводить обучение без предварительной балансировки и фильтрации обучающих данных, а применение рецептивных полей в качестве оптимизируемых параметров решающих правил позволяет не проводить предварительной нормализации (приведение к единой шкале измерения) признаков.

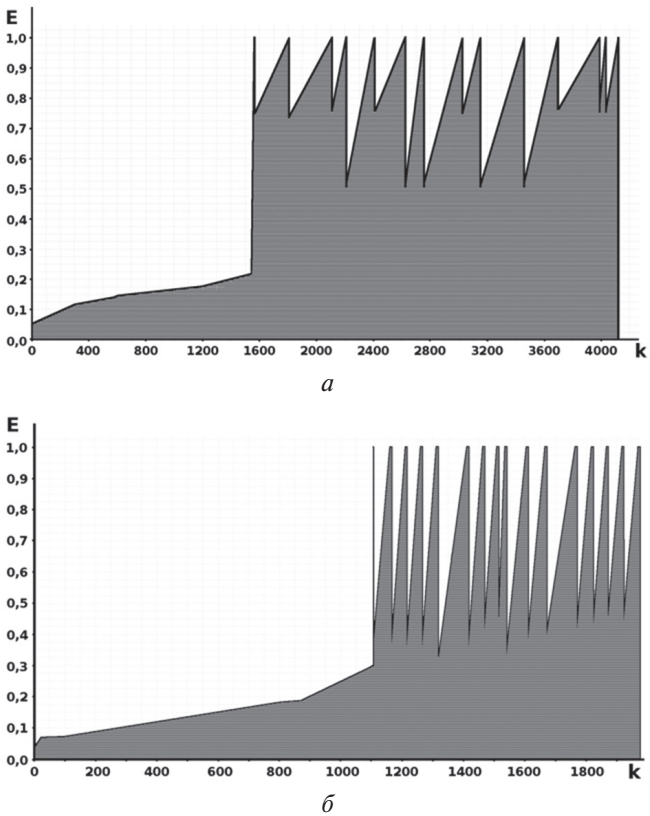
### 3. Результаты физического моделирования

Предложенный алгоритм обучения классификатора по несбалансированным данным большого размера был использован для двух разных практических задач, где было накоплено достаточный объем архивных данных мониторинга управляемого процесса.

Первая задача связанна с распознаванием функционального состояния фронта кристализации при выращивании крупногабаритных монокристаллов. Алфавит из трех классов распознавания характеризует условия выращивания: $X_1^o$ – условия увеличения выпуклости фронта кристализации; $X_2^o$ – оптимальные условия выращивания; $X_3^o$ – условия уменьшения выпуклости фронта кристализации. Словарь из 30 признаков распознавания включает все доступные для контроля параметры, тренды которых архивируются на протяжении каждого выращивания. В первую очередь используются признаки, характеризующие тепловые условия выращивания, а также признаки, характеризующие состояние расплава. В качестве дополнительных признаков распознавания используются разности первого и второго порядка, взятые по трендам основних признаков. Общий обьем архивной истории с метками классов составляет 1 Гбайт данных (4194304 векторов-реализаций). При этом количество реализаций классов, характеризующих отклонение условий выращивания от оптимальних составляет лиш 0,0002% от общего обьема данных.

Вторая задача состоит в распознавании зашифрованного трафика BitTorrent (пиринговые сети обмена файлами) с целью его учета администратором сети при настройке QoS-механизма приоритезации трафика. Наборы данных в процессе трассировки трафика утилитой TcpDump с последующим формированием потоков и вычислением с помощью утилиты NetMate 37-ми признаков распознавания, характеризирующих статистические свойства потоков [2]. Априорная классифика-ция реализаций обучающего трафика основана на результатах мониторинга сокетов утилитой CurrPorts [8]. Общий объем накопленных данных составляет 3,3 Гбайт (13841203), при этом итересующий трафик составляет лишь 0,0001 % от всего объема данных.

Графики изменения максимумов нормированного КФЭ в процессе последовательной оптимизации параметра рецептивных полей в режимах обучения (до первого максимума усредненного КФЭ $E = 1{,}0$) и дообучения (после первого максимума $E = 1{,}0$) показано на рис. 2. Счетчик шагов $k$ соответствует одному изменению параметра какого-либо рецептивного поля.




*а*

*б*

Рис. 2. График изменения максимумов КФЭ в процессе последовательной оптимизации параметра рецептивных полей в режимах обучения и дообучения: *а* – классификатор состояния фронта кристализации; *б* – классификатор трафика BitTorrent

Анализ рис. 2. показывает, что в процессе обучения были получены безошибочне по обучающим случайным выборкам решающие правила.

При этом в процессе дообучения классификатора состояний фронта кристализации было добавлено в обучающую матрицу 13 неверно классифицированных тестовых реализаций. В процессе дообучения классификатора BitTorrent трафика было использовано 15 неверно классифицированных реализаций. После добавления к обучающим выборкам неверно классифицированных реализаций удалось построить безошибочные по обучающим и тестовым матрицам классификаторы.

Таким образом, предложенный алгоритм обучения по несбалансированным выборкам большого размера позволяет осуществить обучение без дополнительной предобработки и построить безошибочные по обучающим и тестовым матрицам классификаторы.

**Выводы**

В рамках информационно-экстремальной технологии синтезировано классификаторы обучающиеся по большим несбалансированным данным, что позволяет осуществить обучение без дополнительной предобработки (нормализации, балансировки, фильтрации), не используя для обучения полудублирующие реализации, что значительно сокращает количество итераций оптимизации параметров.

За результатами физического моделирования по данным архивной базы мониторинга сетевого трафика и процесса кристализации крупногабаритных монокристаллов подтверждено возможность построения безошибочных по обучающим и тестовым выборкам классификаторов при “экстремальной” несбалансированости исходных данных.

**Список літератури: 1.** *Воронцов К.В.* Обзор современных исследований по проблеме качества обучения алгоритмов / К.В. Воронцов // Таврический вестник информатики и математики. – 2004. – №1. – Симферополь, АР. Крым, Украина : КНЦНАУ. – С. 5-22. **2**. *Raman Singh*. Issue related to sampling techniques for network traffic dataset / Raman Sigh, Harish Kumar, R.K. Singla //International Journal of Mobile Network Communications & Telematics. – 2013. – Sydney, Australla : WSP. – Vol.3., No.4. – P. 75-85. **3.** *Yap B.W.* An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets / B.W. Yap, K.A. Rani, H. A. A. Rahman, S. Fong, Z. Khairudin. N.N. Abdullah // Proceedings of the First International Conference on Advanced Data and Information Engineering. Lecture Notes in Electrical Engineering. – 2014. – V.285. – Singapore : Springer Science. – P. 13-22. **4.** *Жора Д.В.* Анализ функционирования классификатора со случайными порогами / Д.В. Жора // Кибернетика и системный анализ. – Киев : Інститут кібернетики ім. В.М. Глушкова НАН України. – 2003. – №3. – С. 72-91. **5.** *Довбиш А.С.* Основи проектування інтелектуальних систем: Навчальний посібник / А.С. Довбиш. – Суми: СумДУ. – 2009. – 171 с. **6.** *Довбиш А.С.* Інтелектуальна система підтримки прийняття рішень для керування вирощуванням монокристалів / А.С. Довбиш, В.С. Суздаль, В.В Москаленко // Вісник СумДУ. Серія технічні науки. – 2011. – №2. – С. 39-47. **7.** *Лэм Ч.* Hadoop в действии / Ч. Лэм. – Москва : ДМК Пресс. – 2012. – 424 с. **8.** *Bujlow T.* Volunteer-Based System for classification of traffic in computer networks / T. Bujlow, K. Balachandran, M.T. Riaz, J.M. Pedersen // In Proceedings of 19th Telecommunications Forum TELFOR 2011’. – 2011. – Sydney : IEEE Press. – P. 210-213.

***Поступила в редколлегию 23.02.2014***

УДК 681.518:004.93.1’

**Інформаційно-екстремальне машинне навчання за незбалансованими даними великого розміру** / В.В. Москаленко // Біоніка інтелекту: наук.-техн. журнал. – 2015. – № 1 (84). – С. 34–38.

Пропонується алгоритм інформаційно-екстремального машинного навчання з ітераційним збором навчальної вибірки, що дозволяє проводити навчання класифікатора на вхідних даних великого обсягу без їх додаткової передобробки (нормалізації, балансування, фільтрації). Як приклад розглянуто реалізацію запропонованого алгоритму для навчання класифікатора по великим незбалансованим наборам архівних даних моніторингу мережевого трафіку і технологічного процесу вирощування сцинтиляційних монокристалів.

Іл.: 2. Бібліогр.: 8 найм.

UDC 681.518:004.93.1’

**Information-extreme machine learning for unbalanced big data** / V.V. Moskalenko // Bionics of Intelligense: Sci. Mag. – 2015. – №1 (84). – P. 34–38.

This article propose an information-extreme learning machine algorithm with an iterative collection of trainig samples that allows learning the classifier on the original large volume of data without preprocessing (normalization, balancing, filtering). Implementation of the proposed algorithm for learning the classifier for large unbalanced set of historical data network traffic monitoring and process of growing single-crystal scintillation are considered.

Fig.: 2. Ref.: 8 items.