DOI: https://doi.org/10.15407/kvt190.04.005

УДК 330.4:004.22

В.І. ГРИЦЕНКО, член-кореспондент НАН України,
директор Міжнародного науково-навчального центру
інформаційних технологій та систем НАН України та МОН України
e-mail: vig@irtc.org.ua

І.М. ОНИЩЕНКО, канд. економ. наук,
старш. наук. співроб. відд. економіко-соціальних
систем та інформаційних технологій
e-mail: standardscoring@gmail.com
Міжнародний науково-навчальний центр
інформаційних технологій та систем НАН України та МОН Украін,
пр. Акад. Глушкова, 40, 03187, м. Київ, Україна

ВИЗНАЧЕННЯ ІНФОРМАТИВНОСТІ ПАРАМЕТРІВ МОДЕЛІ ПРОГНОЗУВАННЯ ЙМОВІРНОСТІ ВИБОРУ ПРОДУКТУ В УМОВАХ «BIG DATA»

Впровадження нових методів та підходів до оброблення даних, які отримали назву «Big Data», особливо актуально для систем з високою завантаженістю. В умовах швидкого потоку даних традиційні пакетні методи моделювання не завжди дають точні та стійкі результати, бракує ефективних методів відбору важливих параметрів. Розглянуто онлайновий підхід до моделювання та прогнозування в умовах «Big Data» та методи оцінювання і відбору параметрів моделі прогнозування ймовірності вибору продукту за їх інформативною важливістю. Для визначення інформативності параметра розглянуто підхід до побудови моделі із використанням регуляризації L1 (LASSO), L2 (RIDGE) та модель Follow-The-Regularized-Leader. Теоретичні та математичні викладки супроводжуються програмною реалізацією методу мовою програмування Python.

Методи online-learning дозволяють отримати оцінки параметрів моделі у режимі реального часу, що дає змогу використовувати їх у високонавантажених системах оброблення даних, у прогнозуванні та прийнятті рішень.

Ключові слова: *інформаційні технології в економіці, економіко-математичне моделювання, алгоритми онлайн навчання, регуляризація, Big Data.*

ВСТУП

Сучасне суспільство переживає чергову хвилю інформаційних технологій, яка цього разу пов’язана зі швидким експоненціальним зростанням обсягів

інформації. При цьому частина структурованої інформації зростає не так стрімко. Основна частина приросту інформації — неструктуровані або слабоструктуровані дані. Класичні ж методи оброблення та зберігання даних не можуть впоратися з такими обсягами та швидкістю приросту даних.

Для розв'язання зазначених вище задач одночасно в кількох найбільших світових компаніях індустрії інформаційних технологій, таких як Google, Facebook та Amazon, почали розробляти абсолютно нові підходи до проблеми зберігання та оброблення інформації з метою отримання корисних знань. Пізніше зусилля окремих компаній було об'єднано у єдиний проект, в результаті якого було отримано систему нових інструментів, методів апаратного та програмного забезпечення для аналізу даних великих обсягів та поганої структурованості. Система таких методів та підходів отримала назву технологій «Big Data».

Термін «Big Data» був введений у 2008 році Кліффордом Лінчем [1], доктором з інформатики Університету Берклі. Також слід відзначити роботи В. Майєр-Шенбергера та К. Кук'єра [2], Ж.-П. Дейкса [3], які проводили фундаментальні дослідження у сфері великих даних.

Хоча сам термін було введено в академічному середовищі, широке застосування та значне поширення він отримав і у практичних дослідженнях в рамках технологічних проектів передових компаній, зокрема у працях Доуг Хеншена (Oracle), Клінта Фінлі (Microsoft), Агама Шаха (Hewlette-Packard), Френкса Білла (Terradata) та інших [4–7].

Останнім часом спостерігається підвищення інтересу до цієї теми з боку дослідників України та країн СНД. Так, Н. Шаховська та Ю. Болюбаш у своїх працях досліджують сучасні бази даних, в яких використано технології «Big Data» [8], Л. Черняк досліджує концепцію «Big Data» в цілому [9], Р. Ускенбаєва досліджує питання впровадження «Big Data» в електронному уряді [10].

У сфері вивчення методів моделювання та оцінювання параметрів моделі із застосуванням «Big Data» слід відмітити праці Р. Беккермана, М. Біленко та Д. Лангфорда [11], Х. Б. МакМахан [12–13].

Треба відзначити, що питання визначення важливості того чи іншого параметра моделі у випадку, коли дані надходять у великій кількості та з великою швидкістю, є недостатньо висвітленим у сучасній науковій літературі.

ПОСТАНОВКА ЗАДАЧІ

Сам термін «Big Data» було введено значно пізніше, ніж практики почали працювати над проблемою великих даних — розроблення методів їх зберігання, оброблення та аналізу. Наразі спостерігається значне підвищення інтересу до цієї галузі як з боку вчених, так і з боку практиків технологічних компаній та бізнесу. Всі ці фактори привели до того, що результатів наукових досліджень з тематики «Big Data» публікується досить багато, включаючи і напрям розроблення ефективних методів моделювання та прогнозування.

У той же час недостатньо висвітленими є практичні питання щодо визначення інформативної важливості параметрів та їх відбору до кінцевої версії моделі, а також питання фільтрації шумів — неважливих параметрів. Також це питання недостатньо вивчено для розвитку технологій «Big Data».

У статті поставлено та розв'язано такі наукові завдання: аналіз застосування алгоритмів онлайн методів моделювання в умовах «Big Data», розроблено комп'ютерну реалізацію методу онлайн навчання з можливістю застосування розріджених векторів початкових даних та L1- і L2- регуляризацій, а також висвітлено питання онлайн методів навчання з фільтрацією параметрів моделі у випадку «Big Data» середовища.

Мета статті — дослідити та модифікувати метод оцінювання і відбору інформативно важливих параметрів за допомогою процедур регуляризації для прогнозування на «Big Data» та здійснити комп'ютерну реалізацію запропонованого алгоритму.

МЕТОДИ ПРОГНОЗУВАННЯ НА «BIG DATA»

Визначення інформативності параметрів моделі завжди було важливою частиною економіко-математичного моделювання. Зазвичай відбираючи параметри для кінцевої моделі, залишають лише найсильніші предиктори та виключають або ігнорують менш важливі.

Але у випадку високонавантажених систем, коли дані надходять з високою швидкістю та великої розмірності — «Big Data», традиційні методи машинного навчання на окремих вибірках підготовлених даних виявляються непродуктивними. Для вирішення проблем ефективного моделювання та прогнозування у системах «Big Data» можна використовувати онлайн методи навчання [14].

Визначення: «Big Data» в інформаційних технологіях — це серія підходів, інструментів та методів оброблення структурованих та неструктурованих даних великих обсягів і різноманітності для отримання результатів, які:

1) легко сприймаються людиною;
2) ефективні в умовах неперервного приросту інформації;
3) дозволяють здійснювати паралельні обчислення, розподілені по численних вузлах обчислювальної мережі.

В якості характеристик, які визначають поняття «Big Data», відзначають «три V»:

1) Volume — об'єм;
2) Velocity — швидкість, як у розумінні швидкості приросту, так і необхідності швидкого оброблення та отримання результату;
3) Variety — різноманітність, у розумінні можливості одночасного оброблення різних типів даних [15].

Традиційно для розв'язання задач класифікації використовують методи машинного навчання, основані на певному фіксованому наборі даних — це так званий пакетний (batch) підхід. При цьому усі дані доступні одразу і можуть бути оброблені на одному обчислювальному вузлі. Також пакетний підхід означає, що модель спочатку була навчена на певному наборі даних — training dataset, а потім тестується на тестовому наборі даних — test dataset, та використовується для прогнозування у практичній діяльності. В основі такого підходу лежить гіпотеза про те, що структура даних та статистичні співвідношення між параметрами моделі не змінюються в часі.

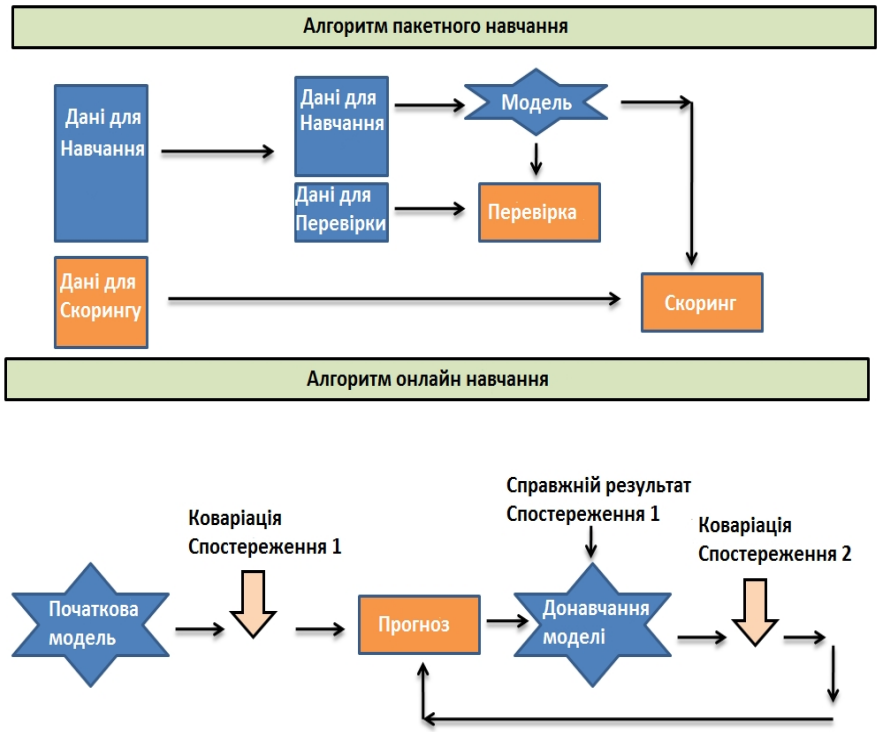

Рис. 1. Порівняння пакетного та онлайнового методів навчання моделі [16].

Спроби розв'язати задачу прогнозування пакетними методами призводили до нестійких у часі результатів, оскільки розмір вибірки для навчання від онлайн аукціону сягає кількох мільйонів записів за один день. Зміна структури вибірки досить відчутна, якщо вивчаються довготривалі процеси з розбиттям по днях тижня. У такому випадку обмежитись набором даних одного дня неприпустимо. Зазначимо, що наявність сильних впливів, наприклад, публікація резонансної новини на сайті, значно змінює тренд переходу за посиланням та саму структуру даних. Збільшення періоду аналізу призводить до значного зростання кількості даних. Але, навіть якщо після оброблення даних отримано надійний прогноз, похибка на наступних кроках прогнозування може швидко зростати через динамічність системи. Це підтверджує нагальну потребу постійного оновлення параметрів моделі для підтримки актуальності та точності прогнозу.

Вказані вище проблеми можна вирішити із застосуванням для моделювання алгоритмів, які дозволяють постійно здійснювати навчання моделі і одночасно отримувати прогнозовані значення (Рис. 1).

Лінійні методи класифікації мають багато переваг для використання у системах «Big Data» завдяки своїй простоті та можливості масштабування і паралельних обчислень. Хоча параметри моделі можуть мати велику розмірність, кількість ненульових коефіцієнтів при параметрах зазвичай складає не більше кількох сотень. Оскільки у процес навчання та роботи моделі залучаються лише параметри з ненульовими коефіцієнтами, це дозволяє економити ресурси та покращити час оброблення даних у порівнянні з іншими методами.

МОДИФІКАЦІЯ FTRL МОДЕЛІ, ОЦІНЮВАННЯ ТА ВІДБІР ПАРАМЕТРІВ

Розглянемо використання алгоритму Follow The Regularized Leader (FTRL) [6] для прогнозування ймовірності вибору продукту за рекламним оголошенням в інтернеті (перехід за посиланням). Цей алгоритм базується на тому, що на кожному кроці вибирається такий набір параметрів об'єкта (в цьому випадку — рекламного оголошення), який приводить до найменшої похибки на цьому кроці:

$$w_t = arg\,min \sum_{i=1}^{t-1} v_i(w) + R(w),$$

де $v_i(w)$ — функція втрат, w — коефіцієнти початкових параметрів моделі, $R(w)$ — функція залишків, t — номер ітерації навчання моделі.

Функція втрат має вигляд:

$$v_t(w) = \|w - x_t\|^2,$$

де x_t — бінарний вектор початкових параметрів моделі, тобто $x_t = 1$, якщо параметр наявний, $x_t = 0$, якщо відсутній.

У випадку лінійної функції оптимізації функція втрат має вигляд:

$$v_t(w) = \langle w, z_t \rangle,$$

де z_t — величина інформативності параметрів на ітерації t.

Оскільки у випадку прогнозування переходу за рекламним посиланням досліджувана величина є бінарною залежною змінною, то зручно використовувати логарифмічну функцію втрат:

$$v_t = (\sigma(w \cdot x_t) - y_t)x_t,$$

де σ — сігмоїдальна функція:

$$\sigma(a) = \frac{1}{1 + e^a},$$

де a — константа, параметр моделі, що відповідає за швидкість навчання.

Використання алгоритму FTRL передбачає обов'язкове залучення повного набору початкових параметрів для прогнозування змін досліджуваної величини. У випадку «Big Data» оброблення повного набору початкових параметрів може призвести до перенавантаження системи та значного зростання вартості розрахунків.

Для оптимізації процесу розрахунків та мінімізації їх вартості запропоновано модифікувати алгоритм FTRL з використанням процедури регуляризації. Такий підхід дозволить відбирати найбільш інформативні параметри для прогнозування з дотриманням належного рівня якості прогнозу. Загальноприйнятими є такі типи регуляризації — L0, L1 та L2 [17].

За результатами попереднього аналізу вибрано функції регуляризації L1 та L2 для зниження розмірності початкового набору параметрів. Тоді функція залишків визначається з урахуванням регуляризації за формулою

$$R(w) = \frac{1}{2\eta}\|w\|^2$$

для деякого $\eta > 0$. Отже, ітерація алгоритму навчання матиме вигляд:

$$w_{t+1} = -\eta\sum_{i=1}^{t} z_i = w_t - \eta z_t \ .$$

Останню рівність можна також писати у вигляді:

$$w_{t+1} = w_t - \eta\nabla v_t(w_t)\ ,$$

що відповідає рівнянню алгоритму покрокового градієнтного спуску.

Остаточно формула обрахунку параметрів моделі FTRL з регуляризацією має вигляд:

$$w_{t,i} = \begin{cases} 0, \text{ якщо } |z_i| \le \lambda_1 \\ -\left(\dfrac{\beta+\sqrt{n_i}}{a} + \lambda_2\right)^{-1} (z_i - sign(z_i)\lambda_1), \text{ якщо } |z_i| > \lambda_1 \end{cases},$$

де a, β — початкові коефіцієнти, які відповідають за швидкість навчання, $\lambda_1, \lambda 2$ — коефіцієнти, які відповідають за силу регуляризацій L1 та L2 відповідно. Вектори z, n розраховуються на кожному кроці (ітерації) разом з коефіцієнтами моделі w та залежать від початкових даних. Формули для розрахунку векторів z, n мають вигляд:

$$v_i = (p_t - y_t)x_i\ ,$$

$$\sigma_i = \frac{1}{a}\left(\sqrt{n_i + v_i^2} - \sqrt{n_i}\right),$$

$$z_i = z_{i-1} + v_i - \sigma_i w_{t,i}\ ,$$

$$n_i = n_{i-1} + v_i^2\ .$$

Таким чином, модель вибору значущих початкових параметрів, яка враховує зазначену вище модифікацію алгоритму FTRL, дозволяє залишати для аналізу тільки ті параметри, які задовольняють умову $w > 0$.

Якщо покласти $\eta = const$, $\lambda_1 = 0$, то отримуємо алгоритм градієнтного спуску. На відміну від алгоритму покрокового градієнтного спуску, за яким на кожному кроці зберігаються коефіцієнти початкових параметрів w, FTRL алгоритм дає змогу зберігати вектор z, а потім на його основі розраховується інформативність початкових параметрів. Таким чином, одночасно виконуються процеси фільтрації неінформативних параметрів та навчання моделі.

Модель FTRL належить до класу «жадібних» алгоритмів (greedy algorithm). Даний клас алгоритмів базується на прийнятті локально оптимального рішення на кожному кроці з метою прийти до глобального оптимуму наприкінці [18, 19]. Метод налаштування та фільтрування коефіцієнтів за допомогою оптимізації з використанням L1-регуляризації отримав назву LASSO (Least Absolute Shrinkage and Selection Operator) [20].

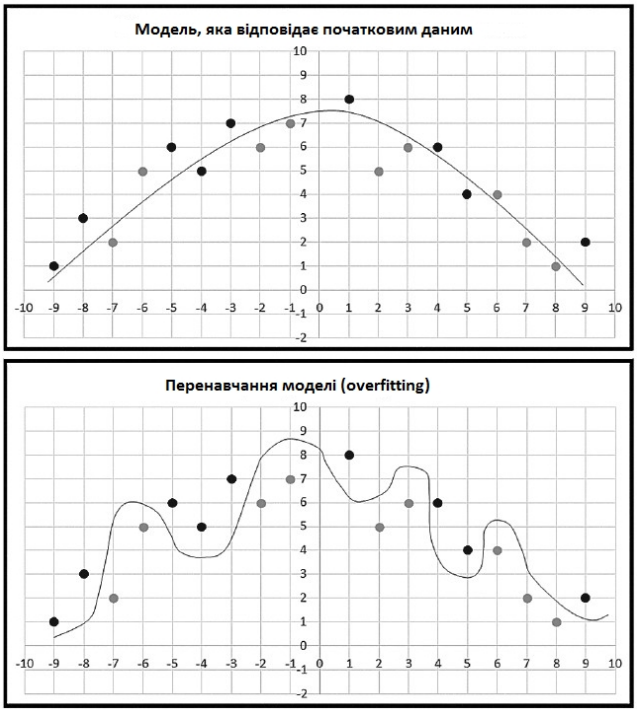

Рис. 2. Перенавчання та нормальна робота моделі [19].

Необхідно відзначити, що використання регуляризацій L1 та L2 дозволяє уникнути такого поняття як перенавчання моделі (overfitting) (Рис. 2).

Отже, регуляризація використовується як «штраф» за велику вагу певного параметра моделі, що дозволяє уникнути перенавчання. Виходячи з доведених властивостей задачі оптимізації під час пошуку рішення з використанням L1-регуляризації, отримано оптимальні рішення, а саме вектор w має властивість розрідженості, тобто частина коефіцієнтів дорівнює нулю.

Таким чином, модифіковано алгоритм FTRL з використанням процедур регуляризації для випадку онлайн навчання, що дозволяє його безпечне використання у «Big Data» системах. Запропоновану модель прогнозування ймовірності вибору продукту розглянуто на прикладі роботи аукціону онлайн реклами.

КОМП'ЮТЕРНА РЕАЛІЗАЦІЯ МОДЕЛІ ПРОГНОЗУВАННЯ ЙМОВІРНОСТІ ВИБОРУ ПРОДУКТУ

Комп'ютерна реалізація запропонованої моделі з використанням мови програмування Python має такий вигляд:

```
    #імпортуємо необхідні бібліотеки
    import numpy as np

    #створюємо клас модель з відповідними функціями та
параметрами
    class FTRLProximal:
```

```
    #функція ініціалізації об'єкту класу
      def __init__(self, n_inputs):
          self.z = np.zeros(n_inputs)
          self.n = np.zeros(n_inputs)

    #сигмоїдна функція
      def sigmoid(self, x):

          return 1 / (1 + np.exp(-x))

    # процедура тренування моделі 1 ітерація з регу-
ляризацією
       def fit_iteration(self, idx, y, alpha, beta,
lambda_1, lambda_2):

           alpha_inv = 1 / alpha

           w = self.weight_update(idx, alpha_inv,
beta, lambda_1, lambda_2)
           p = self.sigmoid(w.sum())
           g = (p - y) #* x_i
           dn = self.n[idx] + np.power(g, 2)
           sigma = alpha_inv * (np.sqrt(dn) -
np.sqrt(self.n[idx]))
           self.z[idx] = self.z[idx] + g -
np.multiply(sigma, w)
           self.n[idx] = dn

           return p, idx, g, w

    # процедура оновлення коефіцієнтів моделі з ре-
гуляризацією
       def weight_update(self, idx, alpha_inv, beta,
lambda_1, lambda_2):
           dw = np.zeros(idx.size)
           mask = np.abs(self.z[idx]) > lambda_1

           z_i = self.z[idx][mask]
           n_i = self.n[idx][mask]

           tmp_1 = z_i - np.sign(z_i) * lambda_1
           tmp_2 = (beta + np.sqrt(n_i)) * alpha_inv +
lambda_2

           dw[mask] = -np.divide(tmp_1, tmp_2)

           return dw
           # процедура для скорингу (прогнозування) з
регуляризацією
```

```
        def predict(self, idx, alpha_inv, beta,
lambda_1, lambda_2):
            w = self.weight_update(idx, alpha_inv,
beta, lambda_1, lambda_2)
            return self.sigmoid(w.sum())
```

Результати комп'ютерних досліджень цієї програмної реалізації на згенерованих даних, які імітують роботу аукціону онлайн реклами, показують, що коефіцієнти параметрів отримують найбільші значення за умови L1 = 0 та монотонно зменшуються зі збільшенням регуляризації (Рис. 3). Тобто за відсутності регуляризації прогнозування здійснюється за усіма початковими параметрами, збільшуючи при цьому ймовірність перенавчання моделі та ускладнюючи розрахунки. Використання регуляризації дозволяє відфільтровувати параметри з низькою інформативністю за умови нульових значень коефіцієнтів. Також слід відзначити, що коефіцієнт регуляризації потрібно обережно підбирати, контролюючи величину похибки прогнозу. У випадку сильної регуляризації можна отримати результат, коли за розробленим алгоритмом всі параметри визначатимуться неінформативними і модель не працюватиме, даючи постійно прогноз, рівним нулю.

Показано, що за алгоритмом LASSO відбираються більш інформативні параметри. У випадку існування значної кореляції між кількома початковими параметрами, в результаті роботи алгоритму LASSO залишається лише один найважливіший параметр, а усі інші — прямують до нуля.

Оскільки за цим методом будується кусково-лінійна траєкторія на просторі початкових параметрів, то крім функції фільтрації, LASSO дозволяє визначити порядок входження параметрів в інформативну множину та виявляти відносну важливість кожного з них.

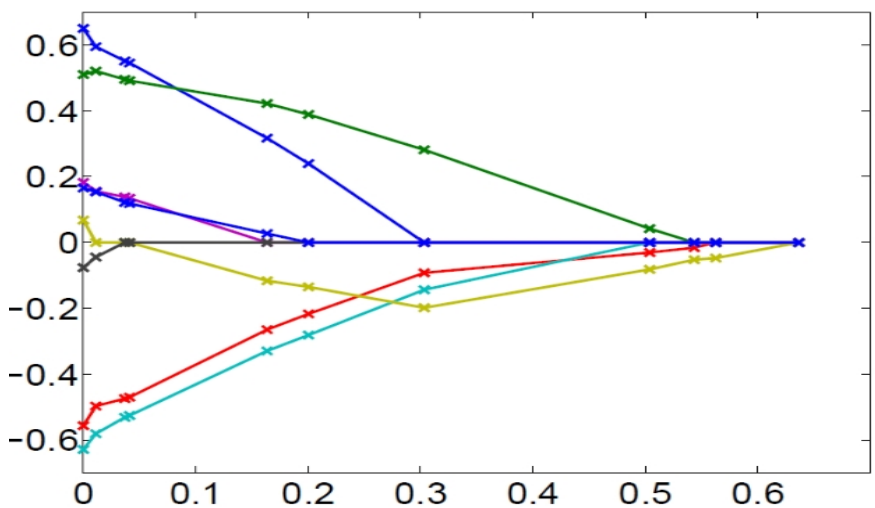

Рис 3. Залежність значення коефіцієнтів параметрів моделі (вісь Y) від коефіцієнта регуляризації (вісь X) [20]

Запропоновану програмну реалізацію моделі FTRL з регуляризацією та механізмом онлайн навчання можна використовувати для прогнозування ймовірності відгуків на рекламні оголошення, активації продукту чи іншої активності в соціальних мережах, онлайн аукціонах, рекомендаційних системах та інших видах інтернет діяльності.

ВИСНОВКИ

Удосконалено модель FTRL на випадок онлайн навчання, що дозволяє ефективно прогнозувати бінарні сигнали при використанні високонавантажених «Big Data» систем. Оскільки кількість параметрів моделі може бути значною, що ускладнює використання моделі та збільшує час і витрати ресурсів на обслуговування процесів моделювання, розв'язати це завдання можна за допомогою використання L1-регуляризації, що дозволяє в режимі реального часу ефективно контролювати кількість параметрів моделі та оцінювати їх відносну інформаційну важливість.

Розроблена програмна реалізація описаної математичної моделі прогнозування ймовірності вибору продукту дозволяє ефективно працювати з розрідженими векторами початкових параметрів та оновлювати лише коефіцієнти тих параметрів, які надано у початковому наборі. Реалізований алгоритм передбачає використання L1- та L2-регуляризацій, що допомогає краще контролювати процес навчання моделі та уникнути її перенавчання.

Запропоновану програмну реалізацію може бути використано для моделювання та прогнозування процесів із швидкими потоками даних, таких як соціальні мережі, онлайн аукціони, ігри, рекомендаційні системи та інші види інтернет діяльності.

ЛІТЕРАТУРА

1. Майер-Шенбергер В. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. Москва, 2014. 240 с.
2. Regelson M., Fain D. Predicting click-through rate using keyword clusters. *Proceedings of the Second Workshop on Sponsored Search Auctions* . Vol. 9623. Citeseer, 2006.
3. M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. Proceedings of the 16th International Conference on World Wide Web. ACM (May 08–12, Banff).Banff, AB, Canada, 2007. P. 521–530.
4. Shalev-Shwartz Shai. Online Learning and Online Convex Optimization. *Foundations and Trends in Machine Learning*. 2011. P. 107–194.
5. Gasso G., Pappaioannou A., Spivak M., Bottou L. Batch and online learning algorithms for nonconvex Neyman-Pearson classification. *ACM Transaction on Intelligent System and Technologies*, 2(3). 2011.
6. H Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. 14th International Conference on Artificial Intelligence and Statistics. (April 11–13, 2011, Ft. Lauderdale). Ft. Lauderdale, FL, USA, 2011. P. 525–533.
7. Фрэнкс Б. Укрощение больших данных: как извлекать знания из массивов информации с помощью глубокой аналітики. Москва, 2014. 352 с.
8. Шаховська Н.Б., Болюбаш Ю.Я. Модель Великих Даних «Сутність — характеристика». URL: http://www.academia.edu/19609620/%D0%9C%D0%9E%D0%94%D0%95%D0% 9B%D0%AC_%D0%92%D0%95%D0%9B%D0%98%D0%9A%D0%98%D0%A5_ %D0%94%D0%90%D0%9D%D0%98%D0%A5_%D0%A1%D0%A3%D0%

A2%D0%9D%D0%86%D0%A1%D0%A2%D0%AC-%D0%A5%D0%90%D0%A0%D0%90% D0 %9A% D0%A2%D0%95%D0%A0%D0%98%D0%A1%D0%A2%D0%98%D0%9A%D0%90_ (11.05.2017)
9. Черняк Л. Большие Данные — новая теория и практика. *Открытые системы. СУБД*. Москва, 2011. № 10. URL: http://www.osp.ru/os/2011/10/13010990/ (11.05.2017)
10. Uskenbaeva, R.K., Kuandykov A.A., Kalizhanova A.U. Tasks of resources provision of distributed computer system's functionality. *World Academy of Science, Engineering and Technology*. 2012. Iss. 70. P. 580–581.
11. R. Bekkerman, M. Bilenko, J. Langford. Scaling up machine learning: Parallel and distributed approaches. Cambridge University Press. 2011.
12. H.B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. 14th International Conference on AISTATS. (April 11–13, 2011, Ft. Lauderdale, FL, USA), Ft. Lauderdale, 2011.
13. H.B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. (June 27–29, 2010, Haifa, Israel), Haifa, 2010.
14. Гриценко В.І., Онищенко І.М. Застосування інструментів Big Data для підвищення ефективності онлайн реклами. Економіко-математичне моделювання соціально-економічних систем. *Збірник наукових праць*. Вип.21. Київ, 2016. С 5–21.
15. Big Data. Wikipedia. URL: https://en.wikipedia.org/wiki/Big_data (11.05.2017)
16. Что такое Real-Time Bidding. URL: http://konverta.ru/how (11.05.2017)
17. Introduction to online machine learning: Simplified. URL: http://www.analyticsvidhya.com/blog/2015/01/introduction-online-machine-learning-simplified-2/ (11.05.2017)
18. Riedman J. H., Hastie T., Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 2010. Vol. 33, № 1. P. 1 – 22.
19. L1- и L2-регуляризация в машинном обучении. URL: https://msdn.microsoft.com/ukua/magazine/dn904675.aspx (11.05.2017)
20. L1-регуляризация линейной регрессии. Регрессия наименьших углов (алгоритм LARS). URL: chrome-extensi-on://ecnphlgnajanjnkcmbpancdjoidceilk/content/web/viewer.html?source=extension_pdfhandler&file=http%3A%2F%2Fwww.machinelearning.ru%2Fwiki%2Fimages%2F7%2F7e%2FVetrovSem11 _LARS.pdf (11.05.2017)

Отримано 28.07.2017

REFERENCES

1. Maier-Shenberher V., Kuker K. Bolshye dannye. Revoliutsyia, kotoraia yzmenyt to, kak my zhyvem, rabotaem y myslym. Moscow, 2014. 240 p. (in Russian).
2. Regelson M., Fain D. Predicting click-through rate using keyword clusters. In Proceedings of the Second Workshop on Sponsored Search Auctions, volume 9623. Citeseer, 2006.
3. Richardson M., Dominowska E., Ragno R. Predicting clicks: estimating the click-through rate for new ads. In Proceedings of the 16th international conference on World Wide Web. P. 521–530. ACM. (May 08-12, 2007, Banff, AB, Canada) Banff, 2007.
4. Shalev-Shwartz Shai. Online Learning and Online Convex Optimization. *Foundations and Trends in Machine Learning.* 2011. P. 107–194.
5. Gasso G, Pappaioannou A., Spivak M., Bottou L. Batch and online learning algorithms for nonconvex Neyman-Pearson classification *ACM Transaction on Intelligent System and Technologies*. 2(3). 2011.
6. H Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. 14th International Conference on Artificial Intelligence and Statistics. .(April 11–13, 2011, Ft. Lauderdale, FL, USA), Ft. Lauderdale, 2011. P. 525–533.
7. Byll Frenks. Ukroshchenye bolshykh dannыkh: kak yzvlekat znanyia yz massyvov ynformatsyy s pomoshchiu hlubokoi analytyky. Moscow, 2014. 352 p. (in Russian)
8. Shakhovska N.B., Boliubash Yu.Ia. Model Velykykh Danykh «Sutnist - kharakterystyka». 2015 r. URL: http://www.academia.edu/19609620/%D0%9C%D0%9E%D0%94%D0%95%D0%9B%D0%AC_%D0%92%D0%95%D0%9B%D0%98%D0%9A

%D0%98%D0%A5_%D0%94%D0%90%D0%9D%D0%98%D0%A5_%D0
%A1%D0%A3%D0%A2%D0%9D%D0%86%D0%A1%D0%A2%D0%AC-%D0%A5
%D0%90%D0%A0%D0%90%D0%A%D0%A2%D0%95%D0%A0%D0%98%D0%
A1 %D0%A2%D0%98%D0%9A%D0%90_ (11.05.2017)
9. Cherniak L. Bolshye Dannye — novaia teoryia y praktyka. Otkrytye systemy. SUBD. Moscow, 2011. № 10. URL: http://www.osp.ru/os/2011/10/13010990/ (11.05.2017)
10. Uskenbaeva, R.K., Kuandykov A.A., Kalizhanova A.U. Tasks of resources provision of distributed computer systems functionalit. *World Academy of Science, Engineering and Technology*. 2012. Iss. 70. P. 580–581.
11. R. Bekkerman, M. Bilenko, and J. Langford. Scaling up machine learning: Parallel and distributed approaches. Cambridge University Press, 2011.
12. H.B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. In AISTATS. (April 11—13, 2011, Ft. Lauderdale, FL, USA), Ft. Lauderdale, 2011.
13. H.B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. In COLT (June 27–29, 2010, Haifa, Israel), Haifa, 2010.
14. Hrytsenko V.I., Onyshchenko I.M. Zastosuvannia instrumentiv Big Data dlia pidvyshchennia efektyvnosti onlain reklamy. Ekonomiko-matematychne modeliuvannia sotsialno-ekonomichnykh system. *Zbirnyk naukovykh prats*. Vyp. 21. Kyiv, 2016. P. 5–21.
15. Big Data. Wikipedia. URL : https://en.wikipedia.org/wiki/Big_data (11.05.2017)
16. Chto takoe Real-Time Bidding. URL: http://konverta.ru/how (11.05.2017)
17. Introduction to online machine learning: Simplified. URL: http://www.analyticsvidhya.com/blog/2015/01/introduction-online-machine-learning-simplified-2/ (11.05.2017)
18. Riedman J. H., Hastie T., Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 2010. Vol. 33, №. 1. P. 1–22
19. L1- y L2-rehuliaryzatsyia v mashynnom obuchenyy. URL: https://msdn.microsoft.com/uk-ua/magazine/dn904675.aspx (11.05.2017)
20. L1-rehuliaryzatsyia lyneinoi rehressyy. Rehressyia naymenshykh uhlov (alhorytm LARS). URL: chrome-extension://ecnphlgnajanjnkcmbpancdjoidceilk/content/web/viewer.html?source=extension_pdfhandler&file=http%3A%2F%2Fwww.machinelearning.ru%2Fwiki%2Fimages%2F7 %2F7e %2FVetrovSem11_LARS.pdf (11.05.2017)

Received 28.07.2017

В.И. Грищенко, член-корреспондент НАН Украины, директор
Международного научно-учебного центра
информационных технологий и систем
НАН Украины и МОН Украины
e-mail: vig@irtc.org.ua.
И.М. Онищенко, канд. эконом. наук,
старш. науч. сотр. отд. экономико-социальных
систем и информационных технологий
e-mail: standardscoring@gmail.com
Международный научно-учебный центр информационных
технологий и систем НАН Украины и МОН Украины,
пр. Акад. Глушкова, 40, 03187, г. Киев, Украина

ОПРЕДЕЛЕНИЕ ИНФОРМАТИВНОСТИ ПАРАМЕТРОВ МОДЕЛИ ПРОГНОЗИРОВАНИЯ ВЕРОЯТНОСТИ ВЫБОРА ПРОДУКТА В УСЛОВИЯХ «BIG DATA»

Внедрение новых методов и подходов к обработке данных, получивших название «Big Data», особенно актуально для систем с высокой загруженностью. В условиях быстрого потока данных традиционные пакетные методы моделирования не всегда дают точные и

устойчивые результаты и не имеют эффективных алгоритмов отбора важных переменных. Рассмотрен онлайновый подход к моделированию и прогнозированию в условиях «Big Data» среды, а также методы оценки и отбора переменых модели по их информативности. Для определения информативности параметра рассмотрен метод построения модели с использованием регуляризаций L1(LASSO) и L2 (RIDGE), а также модель Follow-The-Regularized-Leader. Теоретические и математические результаты сопровождены программной реализацией описанного метода на языке программирования Python.

Методы online-learning позволяют получить оценки информативности параметров модели в режиме реального времени, что дает возможность использовать их для высоконагруженных систем обработки данных, прогнозирования и принятия решений.

***Ключевые слова**: информационные технологии в экономике, экономико-математическое моделирование, алгоритмы онлайн обучения, регуляризация, Big Data.*

V.I. Gritsenko, Corresponding Member of NAS of Ukraine,
Director of International Research and Training
Center for Information Technologies and Systems
of the National Academy of Sciences of Ukraine
and Ministry of Education and Science of Ukraine
e-mail: vig@irtc.org.ua
I.M. Onyshchenko, PhD (Economics), Senior Researcher,
Department of Economic and Social
Systems and Information Technologies
e-mail: standardscoring@gmail.com
International Research and Training Center for Information
Technologies and Systems of the National Academy
of Sciences of Ukraine and Ministry of Education and Science of Ukraine,
40, Acad. Glushkov av., 03187, Kiev, Ukraine

DETERMINING THE INFORMATIVITY OF PARAMETERS
IN A PROGNOSTIC MODEL FOR EVALUATING
THE PROBABILITY OF PRODUCT SELECTION IN CASE OF BIG DATA

Introduction. Fast growth of collected and stored data due to IT bumming caused a problem called "Big Data Problem". Most of the new data are unstructured and this is the core reason why traditional relational data warehouse are so inefficient to deal with Big Data. Predicting and modeling based on Big Data also can be problematic because of high volume and velocity. To avoid some problems online learning algorithms can be successful for high-load systems.

The purpose of the article is to develop an approach to feature selection and modeling in case of Big Data with using online learning algorithm.

Method. Online learning algorithm for FTRL (Follow-The-Regularized-Leader) model with L1 and L2 regularization to select only important features was used.

Results. The approaches of modeling in cases of using batch and online learning algorithms are described on the example of online auction system. The online learning algorithm has very strong preferences in case of high load and high velocity. Mathematical background for modification of linear discriminator of FTL (Follow-The-Leader) model with adding regularization was described. L1 and L2 regularization allows us to select important features in real time. If the feature becomes useless, the regularization will set the corresponding coefficient equal to 0. But it does not remove the feature from training process and the coefficient can be restored with some value in case of its importance for model. The full process is prepared as a program in Python and can be used in practice.

The results may be applied for modeling and forcasting in projects with high volume or velocity of data, for example — social networks, online auctions, online gaming, recommendation systems and others.

Conclusions. FTRL model to work as online learning algorithm that allows to predict binary outcomes in high load Big Data systems was modified.

Getting into account that number of predictors can be enormous it takes much computing resources, time and make the process difficult. This feature selection problem was solved with using L1 regularization. The selection procedure was added to modified online learning FTRL model. L1 regularization to score the importance of predictors in real time was used.

A program that runs described mathematical algorithm was developed. Note that the algorithm effectively works with sparse matrices by analyzing incoming data and updating weights only for predictors that are presented. The algorithm has L1 and L2 regularization features that may be used for feature selection and avoid overfitting.

Keywords: *information technologies in economics, economical and mathematical modeling, online learning algorithms, regularization, Big Data.*