УДК 519.6

ЧЕРТОВ О.Р.

ПОЛІНОМИ КУНЧЕНКА ДЛЯ РОЗПІЗНАВАННЯ ОБРАЗІВ

В статті розглянута задача розпізнавання образів за допомогою застосування поліномів наближення Кунченка. Викладення ведеться на прикладі виділення шаблону (зразка) стереотипної поведінки зловмисника, що займається мереживим шахрайством типу «склікування». Запропоновано новий підхід до аналізу часових рядів, що містять дані про «склікування». Показано, що це дозволяє визначити певні типи недійсних переходів по рекламним посиланням. Обґрунтовано перспективність застосування вказаного підходу до дослідження моделей в області статистики та соціології.

This paper concerns the task of pattern recognition by using Kunchenko's approximating polynomials. We take a stereotyped behavior template matching of intruder during click fraud as an example. New approach for clicks' time series analyzing is proposed. It has been shown, that with the help of proposed techniques it is possible to define some invalid clicks. The prospects for further research in area of statistics and sociology are confirmed.

Вступ

Класифікація об'єктів (образів) за певними категоріями чи класами називається розпізнаванням образів. Задачі, які відносяться до розпізнавання образів, виникають в багатьох областях – від медичної діагностики, статистики, геології до сканування паперових документів чи дактилоскопії. Виділяють чотири групи методів розпізнавання образів [1]:

1) порівняння зі зразком (template matching), коли застосовується геометрична нормалізація та рахується відстань від певного зразка;

2) статистична класифікація (statistical classification), коли для кожного класу будується свій розподіл і здійснюється класифікація за правилом Байєса;

3) синтаксичне чи структурне порівняння (syntactic or structural matching), коли об'єкт ділиться на елементи та класифікується в залежності від того – містить він чи ні окремі елементи або їх послідовності;

4) нейронні мережі (neural networks), коли вибирається певний тип мережі та налагоджуються його коефіцієнти.

В статті ми зупинимося на першій з вказаних груп методів і дослідимо випадок, коли в початковому сигналі необхідно розпізнати заздалегідь відомий зразок (шаблон) [2]. Такі задачі типові при статистичному аналізі даних, в радіолокації, електрозв'язку, сейсморозвідці тощо, коли при моделюванні складної системи довільної природи виникає потреба в поліноміальному наближенні функції.

Наразі найширше застосування при розв'язанні вказаних задач знаходять класичні поліноми Тейлора (чи Маклорена) та поліноми (ряди) Фур'є. Загальновідомо, що поліном Тейлора є поліномом найкращого наближення функції в околі заданої точки. Але функція, що наближується, повинна мати в цьому околі похідні відповідного порядку. Значно слабші умови накладаються при розвиненні функції в ряд Фур'є по системі незалежних ортогональних чи ортонормованих функцій, що утворюють базис відповідного простора зі скалярним добутком. Проте з точки зору розпізнавання в вихідному сигналі якогось певного зразка більш перспективним представляється використання так званих поліномів Кунченка [3]. Пов'язано це з тим, що поліноми Кунченка будуються в особливому підпросторі евклідова чи гільбертова простора, який має породжувальний елемент. В якості такого елементу можна взяти саме шуканий зразок. Таким чином, задача співставлення зі зразком зводиться до більш простої задачі пошуку найближчого (в певній метриці) апроксимаційного полінома Кунченка.

В даній роботі зазначений підхід демонструється шляхом використання поліномів наближення Кунченка, насамперед, для боротьби зі «склікуванням» (click fraud) – одним із різновидів мережевого шахрайства, що має місце в сфері Інтернет-реклами. На даний момент не має одностайного погляду на реальні масштаби цього виду шахрайства, але, наприклад, дослідження спеціалізованої компанії Clickforensics [4] свідчить про те,

що на кінець 2009 р. не менше 15 % всіх переходів за рекламними об'явами в Інтернеті були здійснені або помилково, або з шахрайською метою. Тому пошук нових методів розпізнавання «склікування», зокрема, за допомогою поліномів Кунченка є актуальним.

Огляд останніх досліджень і публікацій

Найбільш повне викладення теорії поліномів Кунченка наведено в [3]. В цій роботі, зокрема, були виділені чотири основні відмінності вказаних просторів та поліномів від інших, які також застосовуються при розв'язанні різних науково-технічних проблем:

1) базис вихідного лінійного простору складається лише з однієї (породжувальної) функції;

2) сам вказаний простір містить лише лінійно незалежні та, в загальному випадку, не ортогональні породжені функції;

3) довільна функція із множини породжених функцій може бути наближена лінійною комбінацією (тобто поліномом, який і називається поліномом Кунченка) із будь-яких інших (додаткових) породжених функцій;

4) оскільки функція, що наближується, та функції, що наближують, є лінійно незалежними, то за будь-якого скінченного порядку полінома наближення не існує таких його коефіцієнтів, при котрих він співпаде з наближуємою функцією.

Але в роботах професора Ю.П. Кунченка та його учнів, зокрема, в [5] і [6], основна увага приділяється застосуванню поліномів наближення для фільтрації негауссівських випадкових процесів та величин в радіотехніці, радіолокації та зв'язку. В статті [7], найбільш близькій за тематикою до даної, будується оптимальний за певним критерієм якості алгоритм, що дозволяє по апріорному моментно-кумулянтному опису сигналу вказати образ, тобто сигнал, якому він відповідає. Але, знову ж таки, фактично розглядається задача розпізнавання постійних радіосигналів на тлі асиметричних та ексцесних негауссівських завад.

В даній роботі ми пропонуємо застосувати поліноми Кунченка в принципово іншій проблемній області, а саме: для розпізнавання в часових рядах, що містять інформацію про переходи користувача за рекламними Інтернет-посиланнями, тих шаблонів (зразків), що, скоріш за все, відповідають поведінці зловмисника, який намагається під час демонстрації контекстної реклами завдати фінансової шкоди.

Контекстна реклама є одним із найбільш ефективних та рентабельних видів реклами, бо тематика рекламних об'яв підбирається у відповідності до прогнозуємих інтересів користувача. Наприклад, об'ява може бути показана разом з результатами пошуку за ключовим словом, що було введене відвідувачем. Таким чином, забезпечується максимально ефективне використання коштів, які вкладаються в рекламну кампанію, оскільки рекламна об'ява демонструється лише потенційним клієнтам. Більше того, за допомогою таргетингу (налагодження показу об'яв за часом та місцерозташуванням відвідувачів) можна ще більше звузити коло клієнтів, відкинувши явно неперспективних.

Проте, суттєвим недоліком контекстної реклами є проблема «склікування» – одного із різновидів мережевого шахрайства, коли імітується «клік», тобто перехід реального користувача за рекламним оголошенням, для зняття з відповідного рекламодавця плати за здійснений перехід. Загалом, всі «кліки» користувача по рекламних оголошеннях можна розділити на дві групи: дійсні, тобто ті, що зроблені користувачем з метою переходу на сайт, який рекламується, і недійсні, тобто такі, що зроблені помилково чи з метою шахрайства.

Існує декілька методів, що наразі застосовуються для боротьби зі «склікуванням».

Найбільш поширений метод полягає в порівнянні поточної активності відвідувачів з їх активністю в минулому. За наявності суттєвих відхилень вважається, що рекламна об'ява «зазнала» нападу зловмисників. Але використовуючи даний метод, неможливо врахувати різке збільшення зацікавленості до пропонуємої продукції. Бо воно може бути викликане не тільки розгортанням запланованої рекламної кампанії, але й зовнішніми обставинами. Наприклад, якщо в якомусь місті трапилася серйозна аварія на електростанції, то з великою ймовірністю зросте попит на продукцію компанії, що займається реалізацією автономних джерел живлення.

Другий метод базується на алгоритмах, що керуються правилами («rules-based algorithms») [8]. Суть методу визначається наяв-

ністю таких певних умов, коли «клік» вважається недійсним. Кожен перехід відвідувача проходить через систему фільтрації цими алгоритмами, в результаті чого приймається рішення про те, чи є відповідний «клік» дійсним або недійсним. Підготовкою таких алгоритмів і їх налагодженням (підбором параметрів) займаються відповідні фахівці.

Третій метод використовує для аналізу передісторію індивідуальної активності користувача. Зі збільшенням накопичуємої бази даних стає простіше визначити дійсні переходи відвідувача, ніж недійсні. Даний метод базується на двох припущеннях: в попередніх моделях поведінки користувача відсутні недійсні «кліки» і ці моделі однозначно визначають майбутню поведінку відвідувачів.

Четвертий метод включає ряд алгоритмів, що базуються на підрахунку та врахуванні кількості показів між послідовними «кліками» [9].

Але всі ці методи фактично є чисто статистичними, в них не враховується поведінка можливого зловмисника.

Постановка задачі

Мета роботи полягає в дослідженні принципової можливості застосування поліномів наближення Кунченка для вирішення задачі співставлення зі зразком на прикладі побудови та аналізу моделей поведінки можливих учасників процесу «склікування», які потенційно можна в подальшому використовувати для визначення частини недійсних «кліків» та їх попередження.

Взаємодія учасників процесу надання контекстної реклами

В процесі створення та проведення рекламної кампанії в Інтернеті, як правило, задіяні наступні учасники: рекламодавець, рекламне агентство, пошукова мережа, рекламний майданчик і клієнт (людина, якій адресовані рекламні повідомлення). Характер їх взаємодії наведено на рис. 1.

Якщо рекламодавець вирішує розмістити свої об’яви в Інтернеті, то він може це зробити самостійно, скориставшись автоматизованими продуктами, котрі надаються, наприклад, такою рекламною мережею як Google. Але в цьому випадку йому потрібно буде входити в особливості налагодження рекламних кампаній.

Альтернативний варіантом є використання послуг рекламних агентств. Перевагою такого підходу є також те, що спеціалізовані агентства мають відповідну кваліфікацію та зможуть забезпечити проведення рекламної кампанії на високому рівні.

Сплата послуг розміщення реклами здійснюється в залежності від обраної моделі рекламної кампанії. Модель CPC (cost per click) передбачає сплачування за кожен перехід за об’явою. Обираючи модель CPM (cost per mille), рекламодавець платить за тисячу показів реклами. Модель CPA (cost per action) гарантує, що оплачуватися буде тільки якась наперед визначена дія відвідувача.

В залежності від налагоджень рекламної кампанії об’ява буде розміщуватися в якості результатів пошуку по ключовим словам, заздалегідь визначеним рекламодавцем, або на сайтах, суміжних за тематичним спрямуванням. Власник сайту, на якому розміщується рекламне повідомлення, отримує певну суму коштів за переходи по об’явам чи за тривалість розміщення останніх на своєму сайті.

Здійснювати певні неправомірні дії, тобто виконувати «склікування» («клікати» по об’явах при моделі CPC) або «споказувати» (генерувати штучні покази при моделі CPM) можуть різні групи учасників. Наприклад, власник сайту – з метою підвищення своїх доходів, рекламодавець – для розтрачання рекламного бюджету своїх конкурентів тощо.

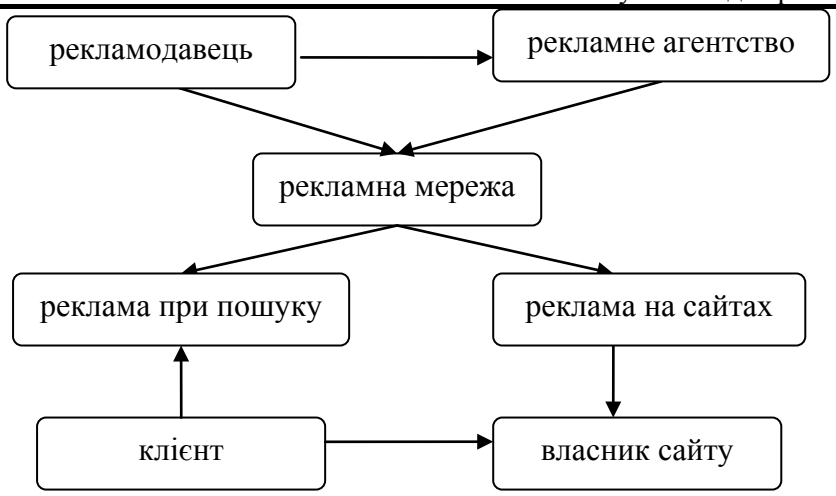

Рис. 1. Взаємодія учасників проблемної області

Часові ряди та шаблони поведінки потенційних зловмисників під час «склікування»

Системи моніторингу Google AdWords/AdSense, Yahoo! Search Marketing чи Microsoft adCenter перевіряють кожен «клік» за такими параметрами як IP-адреса, час виконання тощо. В даній роботі будемо використовувати такі параметри, як кількість переходів за рекламними об'явами на день та щогодини впродовж деякого проміжку часу (тиждень, місяць). Братимемо до уваги лише ті «кліки», що будуть визнаватися дійсними відповідними системами моніторингу.

Розглянемо детально спочатку найбільш розповсюджену ситуацію, коли зловмисник-рекламодавець «склікує» об'яви свого головного конкурента, щоб підірвати його рекламну кампанію. Стратегічна задача зловмисника – «склікати» об'яву таким чином, щоб вона перестала відображатися серед результатів пошуку. Це відбудеться за вичерпання добового бюджету рекламної кампанії (суми, котру рекламодавець згоден витратити на рекламу протягом доби).

Для опису моделі поведінки цього учасника можна скористатися моделлю інформаційної атаки [10]. В нашому випадку інформаційна атака включає такі фази як «фоновий шум», «спроба», «затишшя», «атака». Ці фази відповідають наступним етапам процесу «склікування»:

- природний рівень «кліків», тобто «кліки», що зроблені реальними відвідувачами до тих пір, поки сайт ще не став жертвою «склікування»;
- перші спроби «склікати» об'яву – зловмисник ще не знає напевне, яким чином буде реагувати система моніторингу пошукової мережі, яким є розмір добового бюджету (тобто скільки «кліків» потрібно зробити, щоб об'ява перестала відображатися серед результатів пошуку);
- шахрайський вплив на рекламну об'яву призупиняється на певний час, щоб знизити ймовірність розкриття шахрайства та прийняти рішення про момент нападу;
- власне, атака, в результаті котрої рекламний добовий бюджет об'яви буде витрачений протягом короткого проміжку часу.

Відповідні елементи відобразяться у вигляді локальних та глобальних максимумів на часовому ряді кількості «кліків» (див. рис. 2). Екстремуми відповідають фазам «спроби» і «атаки». При цьому другий максимум буде більшим за перший. Різниця між цими максимумами може бути досить суттєвою, бо в результаті збільшення кількості «кліків» підніметься значення показнику CTR (click-through rate), а, відповідно, вартість одного «кліку» зменшиться. Отже, фазі «атаки» буде відповідати інтервал з найбільш високою щільністю «кліків» на протязі дня.

Як для рекламодавця, що бореться шахрайськими методами зі своїми конкурентами, так і для інших учасників процесу проведення рекламної кампанії в Інтернеті, можні виділити інші стереотипні моделі (шаблони) поведінки.

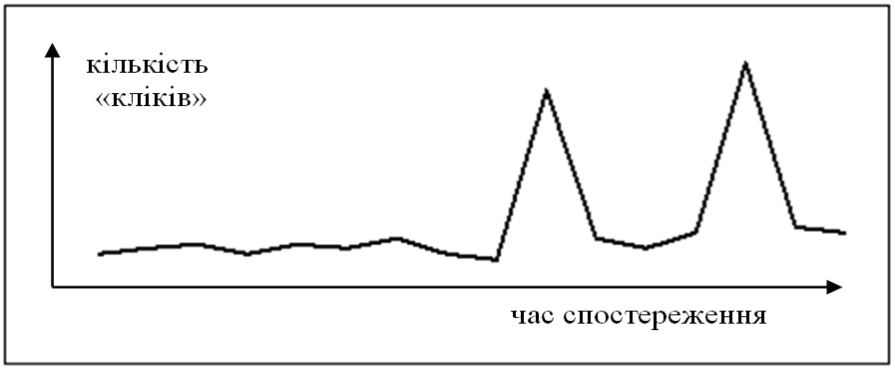

Рис. 2. Модель «склікування» типу інформаційної атаки

Наприклад, «склікувати» рекламні об'яви можуть також зловмисники з метою заподіяння максимальних збитків пошуковим мережам. В цьому випадку дії зловмисників сконцентруються на найдорожчих рекламних оголошеннях. При цьому буде відсутній взаємозв'язок між оголошеннями, а саме: місце розміщення та продукція, що рекламується можуть суттєво відрізнятись. Кількість «кліків» по кожному з оголошень не буде пов'язана жодною залежністю, проте збільшення активності відвідувачів матиме місце в один і той же час. На часовому ряді це матиме вигляд приблизно такий, як на рис. 3.

Рис. 3. Модель поведінки зловмисників при «склікуванні» найдорожчих об'яв

Побудова поліномів Кунченка

Виділені в попередньому розділі (чи аналогічні їм) моделі поведінки зловмисників як певні шаблони (зразки) функціональної залежності кількості «кліків» від часу спостереження можна взяти в якості єдиного базисного елементу вихідного лінійного простору, тобто в якості його породжувального елементу e.

Тоді як лінійну комбінацію лінійно-незалежних перетворень $f_1(e), f_1(e), \ldots, f_n(e)$ відповідного породжувального елементу можна побудувати поліном P_n наближення n–го порядку до (частини) вихідного сигналу $f_s(e)$:

$$P_n = \sum_{\substack{k=0,\\ k\neq s}}^{n} c_k f_k(e),$$

де коефіцієнти c_k знаходяться із умови забезпечення мінімуму відстані між будуємим поліномом та вихідним сигналом. Як показано в [3, с. 95-96] при цьому

$$c_0 = \frac{< f_s(e), f_0(e) > - \sum\limits_{\substack{k=1,\\ k\neq s}}^{n} c_k < f_k(e), f_0(e) >}{< f_0(e), f_0(e) >}$$

а інші коефіцієнти c_k знаходяться як розв'язок системи лінійних рівнянь:

$$\sum_{\substack{k=1,\\ k\neq s}}^{n} c_k F_{i,k} = F_{i,s},\ \overline{i=1,n},\ i \neq s,$$

де центровані корелянти $F_{i,k}$ також рахуються за допомогою скалярних добутків відповідних перетворень:

$$F_{i,k} \equiv < f_i(e), f_k(e) > - \frac{< f_i(e), f_0(e) > \cdot < f_k(e), f_0(e) >}{< f_0(e), f_0(e) >}.$$

Чисельною характеристикою, яку можна застосовувати в критеріях якості співставлення сигналу з виділеним шаблоном, тобто як міру наближення полінома Кунченка P_n до (частини) вихідного сигналу $f_s(e)$ є коефіцієнт ефективності d_n:

$$d_n = \frac{\sum\limits_{\substack{k=1,\\ k\neq s}}^{n} c_k < f_k(e), f_s(e) >}{< f_s(e), f_s(e) >}.$$

Застосування поліномів Кунченка при дослідженні моделей в статистиці та соціології

Розглянутий метод розпізнавання певних зразків за допомогою побудови простору з породжувальним елементом та пошуку коефіцієнтів відповідного полінома Кунченка може бути використаний в будь-якій проблемній області, в якій можна апріорі в часовому чи просторовому ряді виділити певні характерні шаблони. Особливо перспективним представляється його застосування до так званих «м'яких» моделей (в термінології Рене Тома). При побудові таких моделей виходять із припущень та гіпотез про суть явищ та процесів, які описуються, роблять висновки з цих гіпотез та уточнюють самі гіпотези. «М'які» моделі характерні, насамперед, для статистики і соціології.

Наприклад, маючи такий інструментарій як інформаційно-аналітична система обробки даних Всеукрїнського перепису населення 2001 р. [11], можна досить просто відшукати певні закономірності розподілу респондентів за територією, сімейним станом, освітою тощо. Перевірку ж того, наскільки віднайдена гіпотеза справедлива для генеральної сукупності, можна здійснити за допомогою описаного метода.

Висновки

Побудувавши типові моделі поведінки можливих мережевих зловмисників під час надання контекстної реклами і створивши шаблони на їх основі, можна застосовувати метод на базі поліномів Кунченка для визначення (і попередження) можливого нападу. Для цього потрібно в якості породжувального елементу відповідного лінійного простору взяти виділений шаблон поведінки зловмисника та, скориставшись коефіцієнтами ефективності, виділити в часовому ряді кількості «кліків» даний шаблон як зразок (в розумінні теорії розпізнавання образів).

В роботі також обґрунтована перспективність застосування вказаного підходу до дослідження моделей в області статистики та соціології.

Список посилань

1. Jain A.K., Duin P.P.W., Mao J. Statistical Pattern Recognition: A Review // IEEE Transactions on pattern analysis and machine intelligence. – 2000. – Vol. 22, № 1. – P. 4-37.
2. Brunelli R. Template matching techniques in computer vision: theory and practice. – Chippenham: Wiley, 2009. – 348 p.
3. Кунченко Ю.П. Приближения в пространстве с порождающим элементом. – К.: Наук. думка, 2003. – 243 с.
4. Click Fraud Index [Electronic resource]. – Режим доступу: http://www.clickforensics.com/resources/click-fraud-index.html
5. Кунченко Ю.П., Заболотній С.В. Поліноміальні оцінки параметрів близьких до гауссівських випадкових величин. Частина ІІ. Оцінка параметрів близьких до гауссівських випадкових величин. – Черкаси: ЧІТІ, 2001. – 251 с.
6. Лега Ю.Г., Гончаров Ю.Г., Філіпов В.В. Асимпотичні властивості оцінок параметра постійного сигналу при усіченому оцінюванні дисперсії асиметричної завади другого типу першого виду // Вісник Черкаського державного технологічного університету. – 2008. – № 3.– С. 3-8.
7. Палагін В.В., Жила О.М. Поліноміальне вирішення задач розпізнавання випадкових сигналів // Вісник Черкаського державного технологічного університету. – 2008. – № 2. – С. 31-35.
8. Matin S. Click Ahoy! Navigation Online Advertising in a Sea of Fraudulent Clicks // Berkeley Technology Low Journal, Annual Review 2007. – Vol. 22, № 1. – P. 533-554.
9. Immorlica N., Jain K., Mahdian M., Talwar K. Click Fraud Resistant Methods for Learning Click-Through Rates // Technical Report. – 2005. – Vol. 3828. – P. 34-45.
10. Фурашев В.М., Ланде Д.В. Практичні засади прогнозування можливих загроз та ризиків шляхом аналізу взаємозв'язку подій з інформаційним простором // Открытые информационные и компьютерные интегрированные технологии: сб. науч. трудов. Вып. 42. – Харьков: Нац. аэрокосм. ин-т «ХАИ», 2009. – С. 194-203.
11. Чертов О.Р. Система многомерного анализа данных Всеукраинской переписи населения 2001 года // Россияне в зеркале статистики: Всероссийская перепись населения 2002 года: Международный симпозиум, 30-31 марта 2004 г.: труды симп. – М.: Изд-во Федеральной службы государственной статистики, 2004. – С. 234-238.

Поступила в редакцию 4.12.2009