

Корначевський Я.І.

ННК “ІПСА” НТУУ “КПІ”

Системи Data Mining

Одним із результатів життєдіяльності людини є накопичення різноманітних даних (інформації). Дані генеруються всіма галузями – від сільського господарства і видобувної промисловості до генетики і астрофізики. Причому, зовсім “просто” з точки зору науки галузі часто генерують значні об’єми даних. Наприклад, урожайність зернових на різних ділянках господарств чи варіації концентрації корисної породи в різних пластах залягання. І кількість цих даних зростає в геометричній прогресії. Постає запитання: як їх використовувати? Адже опрацювання значних масивів даних вимагає особистого підходу в кожному конкретному випадку, оскільки необхідно враховувати чинники, характерні тільки для цієї предметної галузі. Фахівців, які мали б знання, наприклад, у сільському господарстві і одночасно в автоматизованій (або навіть автоматичній) інтелектуальній обробці великих масивів даних, дуже мало або немає зовсім. Один з виходів – розробка систем, які надають готовий математичний апарат аналізу даних – системи *Data Mining*.

Головне, що повинна зробити система Data Mining, – це віднайти *приховану закономірність* і показати її кінцевому користувачеві. Для цього потрібно накопичити дані і виконати процес дослідження – “провести розкопки”.

Виділяють п’ять основних типів закономірностей [1]:

1) *асоціація* – кілька подій пов’язані, наприклад, люди часто купують одночасно порошок для прання і засоби для відбілювання. Знаючи особливості цієї закономірності, можна спрогнозувати необхідний об’єм закупівлі і ввести стимулюючу знижку, що в результаті дасть значний економічний ефект;

2) *послідовність* – події пов’язані одна з одною в часі, наприклад, після купівлі туристичної путівки в Єгипет турист (вже прибувши до курорту) майже завжди купує тур до Каїру і далі (з меншою ймовірністю) до Луксору або коралових островів. Тут можна теж запропонувати гнучку систему знижок, ефективно конкуруючи з місцевими туроператорами і отримуючи додатковий прибуток;

3) *класифікація* – виявляються ознаки, які характеризують групу, до якої належить об’єкт – аналізуються класифіковані об’єкти і для них формується певний набір правил. Приклад – задача розпізнавання;

4) *кластеризація* – досліджувані об’єкти розділяється на однорідні групи. Приклад – задача виявлення стійкості різних сортів пшениці до борошністої роси;

5) *прогнозування* – виявлення динаміки системи на основі накопичених (раніше) даних, наприклад, швидкість мутації вірусів, прогнозування набуття ними стійкості до нових ліків.

В цілому системи Data Mining зводяться до двох основних типів: базованих на основі нейронних мереж і системи логічного аналізу. Перші відносно легко обробляють значні обсяги даних, але не дають відповідь на запитання “чому” – не пояснюють свої результати. Другі для своєї роботи вимагають значних обчислювальних ресурсів, оскільки зводяться до перебору можливих варіантів. Тому однозначною вимогою до таких систем є можливість їх роботи на кластерних архітектурах, навіть за умови застосування гнучких алгоритмів обмеження ділянки пошуку (наприклад WizWhy [2]).

Зважаючи на значну вартість комерційних систем Data Mining, пропонується обрати кілька безкоштовних реалізацій і дослідити їх переваги та недоліки на встановленому в НТУУ “КПІ” кластері. До розгляду пропонуються системи, перелічені в Open Directory project [3].

Література

1. Чубукова І.А. Data Mining, 2-е изд., испр. – Москва: Интернет-Университет Информационных Технологий; БИНОМ., 2008. – 382 с.
2. Система WizWhy, <http://www.wizsoft.com>.
3. Open Directory project, http://www.dmoz.org/Computers/Software/Databases/Data_Mining/Public_Domain_Software.