УДК 378:004.384

ТЕХНОЛОГІЇ DATA MINING ТА MACHINE LEARNING

В. В. Слєсарєв, В.Ю. Михалик
(Україна, Дніпро, НТУ «Дніпровська політехніка»)

DataMining - дослідження і виявлення "машиною" (алгоритмами, засобами штучного інтелекту) в сирих даних прихованих знань, які раніше не були відомі, нетривіальні, практично корисні, доступні для інтерпретації людиною.

Основними завданнями аналізу даних є:

- Завдання класифікації зводиться до визначення класу об'єкта по його характеристикам. Необхідно зауважити, що в цьому завданні безліч класів, до яких може бути віднесений об'єкт, відомо заздалегідь.
- Завдання регресії подібно задачі класифікації дозволяє визначити за відомими характеристиками об'єкта значення деякого його параметра. На відміну від завдання класифікації значенням параметра є не кінцеве безліч класів, а множина дійсних чисел.
- При пошуку асоціативних правил метою є знаходження частих залежностей (або асоціацій) між об'єктами або подіями. Знайдені залежності представляються у вигляді правил і можуть бути використані як для кращого розуміння природи аналізованих даних, так і для передбачення появи подій.
- Завдання кластеризації полягає в пошуку незалежних груп (кластерів) і їх характеристик у безлічі аналізованих даних. Вирішення цього завдання допомагає краще зрозуміти дані. Крім того, угрупування однорідних об'єктів дозволяє скоротити їх число, а отже, і полегшити аналіз.

DataMining може складатися з двох або трьох стадій:

Стадія 1. Виявлення закономірностей (вільний пошук).

Стадія 2. Використовування виявлених закономірностей для прогнозу невідомих значень (предикатив немоделювання). На додаток до цих стадій іноді вводять стадію оцінювання (валідації) , наступну за стадією вільного пошуку. Мета валідації- перевірка достовірності знайдених закономірностей.

Стадія 3. Аналіз виключень - стадія призначена для виявлення і пояснення аномалій, знайдених в закономірностях.

Мета технології DataMining - знаходження в даних таких моделей, які не можуть бути знайдені звичайними методами.

На сьогоднішній день існує безліч алгоритмів машинного навчання, на основі яких можна побудувати модель (представлені на мал. 1): DecisionTree (дерево прийняття рішень), KNN (метод k-найближчих сусідів), SVM (метод опорних векторів), NN (нейромережа). І вибір моделі варто засновувати на те, чого ми від неї хочемо. По-перше, наскільки рішення, що вплинули на результати моделі, повинні бути зрозумілими. По-друге, можливість змінювати пристрій

моделі від одношарової нейронної мережі, до багатошарової, що володіє відмінною здатністю знаходити нелінійні залежності, змінюючи при цьому всього пару рядків коду.

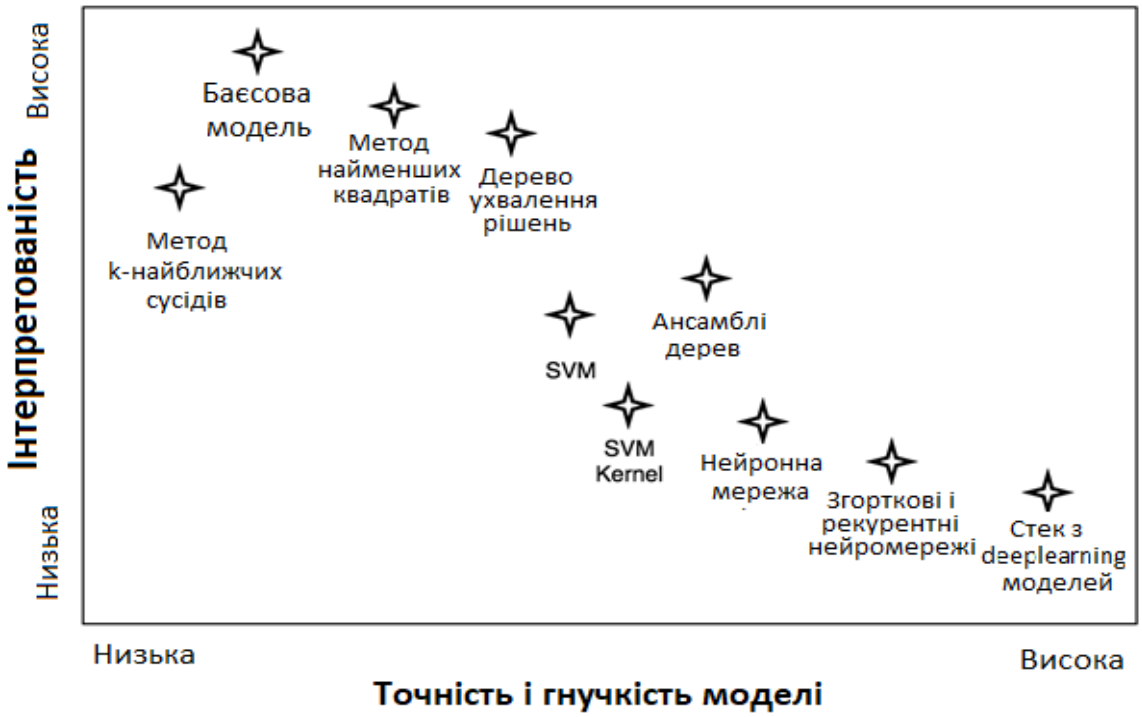

Рис. 1. Порівняння алгоритмів машинного навчання

Класифікація - це найпростіша і поширена задача DataMining. В результаті рішення задачі класифікації виявляються ознаки, які характеризують групи об'єктів досліджуваного набору даних - класи; по цих ознаках новий об'єкт можна віднести до того або іншого класу.

У 1943 році Уоррен Мак-Каллок і Уолтер Піттс запропонували модель математичного нейрона, а в 1958 році Френк Розенблат на основі нейрона Мак-Каллока-Питтса створив комп'ютерну програму, а потім і фізичний пристрій - перцептрон. З цього і почалася історія штучних нейронних мереж. Розглянемо структурну модель нейрона на рис. 2.

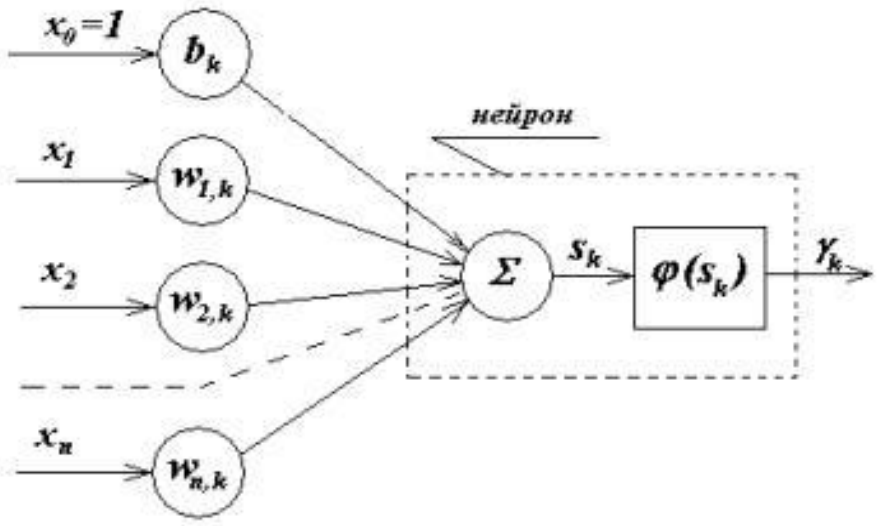

Рис. 2 Функціональна схема моделі штучного нейрона

де:

1. X - вхідний вектор параметрів. Вектор (стовпець) чисел (біол. Ступінь активації різних рецепторів), які прийшли на вхід нейрона. W - вектор ваг (в загальному випадку - матриця ваг), числові значення, які змінюються в процесі навчання (біол. навчання на основі синаптичної пластичності, нейрон вчиться правильно реагувати на сигнали з його рецепторів).

2. Суматор - функціональний блок нейрона, який складає все вхідні параметри помножені на відповідні їм ваги.

3. Функція активації нейрона - є залежність значення виходу нейрона від значення прийшов від суматора.

4. Наступні нейрони, куди на один з безлічі їх власних входів подається значення з виходу даного нейрона (цей шар може бути відсутнім, якщо цей нейрон останній, термінальний).

Потім з цих мінімальних структурних одиниць збирають класичні штучні нейронні мережі. Топологія мережі складається з трьох шарів:

- Вхідний (рецепторний) шар - це вектор параметрів (ознак). Цей шар не складається з нейронів. Можна сказати, що це цифрова інформація, знята рецепторами з «зовнішнього» світу. Прошарок повинен містити стільки елементів, скільки вхідних параметрів (плюс bias-term, потрібний для зсуву порогу активації).

- Асоціативний (прихований) шар - глибинна структура, здатна до запам'ятовування прикладів, знаходженню складних кореляцій і нелінійних залежностей, до побудови абстракцій та узагальнень. У загальному випадку це навіть не шар, а безліч шарів між вхідними та вихідними. Можна сказати, що кожен шар готує новий (більш високорівневий) вектор ознак для наступного

шару. Саме цей шар відповідає за появу в процесі навчання високорівневих абстракцій.

- Вихідний шар - це шар, кожен нейрон якого відповідає за конкретний клас. Вихід цього шару можна інтерпретувати як функцію розподілу ймовірності приналежності об'єкта різних класів. Шар містить лише один нейрон класів представлено в навчальній вибірці. Якщо класу два, то можна використовувати два вихідних нейрона або обмежитися лише одним. В такому випадку один нейрон як і раніше відповідає тільки за один клас, але якщо він видає значення близькі до нуля, то елемент вибірки по його логіці повинен належати іншого класу.

На вході нейрона маємо вектор параметрів, представлені в числовій формі X (i) = {x (i) 1, x (i) 2, ..., x (i) n}. При цьому кожного x(i) зіставлений Y (i) - клас, що задовільнює нашій умові. Нейромережа, по суті, повинна знайти оптимальну, що розділяє гіперповерхность в векторному просторі, розмірність якого відповідає кількості ознак. Навчання нейронної мережі в такому випадку - знаходження таких значень (коефіцієнтів) матриці ваг W, при яких нейрон, який відповідає за клас, буде видавати значення близькі до одиниці в тих випадках, близьких до виконання умови, і значення близькі до нуля, якщо далеко до виконання умови.

$$h_w(X) = f(\sum_{k=1}^{|w|} w_k x_k) \equiv \sigma(w \cdot x)$$

Як видно з формули, результат роботи нейрона - це функція активації (часто позначається черезf(x)) від суми твори вхідних параметрів на шукані в процесі навчання коефіцієнти.

Коли є і навчальна вибірка, і теоретичні знання, можна починати навчання моделі. Однак проблема полягає в тому, що часто елементи множин представлені в нерівних пропорціях. Напевно, метрика точності в такому випадку повинна бути іншою, та й навчати потрібно теж розумно, щоб нейромережа не зробила такого ж очевидно невірного висновку. Для цього можна «годувати» нейромережу навчальними прикладами, що містять однакову кількість елементів різних класів. Вибравши модель і алгоритм навчання, бажано розділити вибірку на частини: провести навчання на навчальній вибірці, що становить 70% від всієї, і пожертвувати 30%.

Підготувавши модель, необхідно адекватно оцінити її якість. Для цього і снують такі поняття:

- TP (TruePositive) – істино позитивний. Класифікатор вирішив, результат має бути позитивним і він був позитивним.

- FP (FalsePositive) - хибнопозитивний. Класифікатор вирішив, що результат має бути позитивним, але він був негативним. Це так звана помилка першого роду. Вона не така страшна, як помилка другого роду, особливо в тих випадках, коли класифікатор - тест на якесь захворювання.
- FN (FalseNegative) - псевдонегативну. Класифікатор вирішив, результат має бути негативним, але він міг бути позитивним (чи був позитивним). Це так звана помилка другого роду. Зазвичай при створенні моделі бажано мінімізувати помилку другого, навіть збільшивши тим самим помилку першого роду.
- TN (TrueNegative) – істинонегативний. Класифікатор вирішив, що результат має бути негативним і він був негативним.

Таким чином, найпростіша метрика - це метрика достовірності (англ. Accuracy). Але ця метрика не повинна бути єдиною метрикою моделі, як вже зрозуміли. Особливо в тих випадках, коли існує перекіс у вибірці, тобто представники різних класів зустрічаються з різною ймовірністю.

$$Accuracy = \frac{tp+tn}{tp+fp+fn+tn}$$

ПЕРЕЛІК ПОСИЛАНЬ:

1. Барсегян, А. А. Анализданных и процессов: учеб. пособие//БХВ-Петербург, 2009. – СПб, 2009 – 3-е изд. – с. 68-81
2. Інтернет- ресурс: https://habrahabr.ru/post/340792/