УДК 681

Використання нейромереж у процесі інтелектуального (кластерного) аналізу даних

Лещинський О.Л.
кандидат фізико-математичних наук,
доцент кафедри економічної кібернетики
Національного авіаційного університету

Іщенко А.О.
магістр
Національного авіаційного університету

У статті розглянуто проблеми використання нейронних мереж у процесі інтелектуального аналізу даних як інструменту для кластеризації даних. Проведено порівняння алгоритмів, що найбільше підходять для використання у завданнях кластеризації, та подано рекомендації щодо вибору оптимального алгоритму.

Ключові слова: інтелектуальній аналіз даних, кластеризація, нейронна мережа, мережа Кохонена, алгоритм.

Лещинский О.Л., Ищенко А.А. ИСПОЛЬЗОВАНИЕ НЕЙРОСЕТЕЙ В ПРОЦЕССЕ ИНТЕЛЛЕКТУАЛЬНОГО (КЛАСТЕРОГО) АНАЛИЗА ДАННЫХ

В статье рассмотрены особенности использования нейронных сетей в процессе интеллектуального анализа данных как инструмента для кластеризации данных. Проведено сравнение алгоритмов, наиболее подходящих для использования в задачах кластеризации, и даны рекомендации по выбору оптимального алгоритма.

Ключевые слова: интеллектуальной анализ данных, кластеризация, нейронная сеть, сеть Кохонена, алгоритм.

Leschynsky O.L., Ishchenko A.O. USING OF NEURAL NETWORKS IN THE DATA MINING (CLUSTERING) PROCESS

The article considers the peculiarities of neural networks in the process of data mining, as an instrument for data clustering, comparison of different algorithms that are best suited for using in clustering problems and provides guidance on the selection of the algorithm.

Keywords: data mining, clustering, neural network, Kohonen network, algorithm.

Постановка проблеми у загальному вигляді. У сучасному світі в результаті постійно зростаючого інформаційного потоку виникає проблема його оброблення та виявлення прихованих закономірностей. Такі закономірності відіграють важливу роль в оцінюванні стратегії та тактики будь-якої установи та виявленні її потенціалу.

Аналіз останніх досліджень і публікацій. Вагомий вклад у розвиток сучасного інтелектуального аналізу даних та дослідження властивостей нейронних мереж зробили такі науковці, як В.Ає Дюк [1], І.Д. Мандель [4], А.О. Стариков [5], Н.Г. Ярушкіна [7].

Виділення невирішених раніше частин загальної проблеми. Проблеми використання кластеризації посідають важливе місце в аналізі даних, оскільки результати кластерного аналізу значно впливають на формування стратегії дій підприємств. Відомо, що не існує єдиного правильного алгоритму кластеризації.

Під час використання будь-якого алгоритму кластеризації важливо проаналізувати його позитивні та негативні аспекти, обрати найбільш прийнятні алгоритми та оцінити роль нейронної мережі як інструменту для кластерного аналізу даних.

Формулювання цілей статті (постановка завдання). Незважаючи на досягнення сучасної статистики, у результаті швидкого розвитку комп'ютерних технологій та науки про бази даних обсяг інформації невпинно росте. Сучасні статистичні методи вже не здатні адекватно опрацьовувати великі масиви даних. Інтелектуальний аналіз даних дає можливість виявляти приховані зв'язки у великих масивах інформації.

Метою статті є оцінка нейронних мереж як інструменту для кластерного аналізу даних.

Виклад основного матеріалу дослідження. У результаті розвитку систем управління базами даних та технологій баз даних відбувається значний ріст обсягу даних, що

зберігаються у базах. Цей обсяг даних зазвичай має значний інформаційний потенціал, який може бути розкритий за допомогою технологій інтелектуального аналізу даних, що дає змогу обробляти великі масиви інформації та виявляти в них латентні правила і закономірності [1].

Протягом тривалого часу основою інтелектуального аналізу даних була звичайна математична статистика, яка зазвичай є корисною в умовах перевірки вже сформованих гіпотез [1].

Інтелектуальний аналіз даних являє собою процес виявлення придатних до використання відомостей у великих наборах даних. В інтелектуальному аналізі даних застосовується математичний аналіз для виявлення закономірностей і тенденцій, що існують у даних. Зазвичай такі закономірності не можна виявити під час традиційного вивчення даних, оскільки зв'язки занадто складні або обсяг даних є надмірним.

На початку свого розвитку використання нейронних мереж в аналізі даних викликало неоднозначні відгуки через такі їхні недоліки, як складність структури, занадто довгий період навчання та погана інтерпретованість. Але вони були скомпенсовані комплексом позитивних якостей, таких як низький коефіцієнт помилок, постійне покращення та оптимізація різноманітних алгоритмів навчання мереж, алгоритму отримання правил, алгоритму спрощення мереж, що роблять нейронні мережі надзвичайно перспективним напрямом у сфері аналізу даних [2].

Сферами використання нейронних мереж є прогнозування, класифікація, кластеризація, адаптивне управління, створення експертних систем, автоматизація процесів розпізнавання зображень, обробка аналогових та цифрових сигналів, синтез та ідентифікація електронних ланцюгів і систем тощо [3].

Кластеризація, або природна класифікація – це процес об'єднання у групи об'єктів, що мають схожі ознаки. На відміну від звичайної класифікації, де кількість груп об'єктів фіксована, тут ні групи, ні їх кількість заздалегідь не визначені і формуються у процесі роботи системи, виходячи із близькості об'єктів.

Кластеризація застосовується для вирішення багатьох прикладних завдань – від сегментації зображень до економічного прогнозування та боротьби з електронним шахрайством.

Завдання кластеризації є актуальним, оскільки зростаюче накопичення обсягу даних приводить до необхідності їх класифікації. Під час аналізу об'єктів або явищ стає необхідним врахування все більшої кількості параметрів, тому постає завдання розроблення і застосовування методів, які спеціалізуються на класифікації багатовимірних даних.

Комп'ютерні технології автоматичного інтелектуального аналізу даних переживають бурхливий розквіт. Це пов'язано з потоком нових ідей, що виходять з області комп'ютерних наук, які сформувалися на перетині штучного інтелекту, статистики та теорії баз даних. Елементи автоматичної обробки і аналізу даних стають невід'ємною частиною концепції електронних сховищ даних і часто іменуються як data mining (видобування знань із даних).

Часто виникає необхідність якимось чином класифікувати дані або знайти в них закономірності. Цього можна домогтися, використовуючи як алгоритми кластеризації і методи нейронних мереж, так і методи обробки нечітких мереж [3, 4, 5].

Моделі нейронних мереж можна умовно поділити на три типи, такі як:

1) мережі прямого поширення – одна з найбільш поширених архітектур, яка використовується в прогнозуванні і розпізнаванні образів;

2) мережі зі зворотним зв'язком, які використовуються для оптимізації обчислень та асоціативної пам'яті;

3) самоорганізовувальні мережі, що містять моделі адаптивної резонансної теорії і моделі Кохонена та використовуються для кластерного аналізу.

Останнім часом ведуться досить активні розроблення алгоритмів кластеризації, які здатні обробити дуже великі бази даних. Саме в них основна увага приділяється масштабованості. Розроблено алгоритми, де методи ієрархічної кластеризації інтегровані з іншими методами. До найбільш актуальних алгоритмів відносять BIRCH, CURE, ROCK, Хамелеон, Кохонен [6]. Порівняння цих методів представлено в таблиці, де знаками + та – позначено наявність чи відсутність певної характеристики, а значення +– вказує на те, що в деяких ситуаціях характеристика присутня, а в деяких – ні, що може залежати від вибірки, налаштувань алгоритму чи інших можливих факторів.

У таблиці 1 продемонстровано порівняння алгоритмів кластеризації.

На основі наявного стану розвитку методів та алгоритмів кластеризації і аналізу вхідних даних можна дійти висновку, що реальні

дані дуже відрізняються за характеристиками досліджуваної вибірки, тому для оптимального аналізу доцільно обробляти різні вибірки різними методами.

Із таблиці видно, що штучні нейронні мережі, а саме карти Кохонена, є серйозним конкурентом алгоритмам, що побудовані на графах.

Відома велика кількість типів аналізу даних, заснованих винятково на нейронних мережах, але виділяють два з них, що є найбільш популярними. Вони засновані на самоорганізаційних нейронних мережах і на нечітких мережах.

1. Аналіз даних, оснований на самоорганізаційній нейронній мережі. Самоорганізаційний процес – процес навчання без учителя. За такого навчання [7] навчальна множина складається зі значень вхідних змінних, а у процесі навчання немає порівнювання виходів нейронів із бажаними значеннями. Можна сказати, що така мережа вчиться розуміти структуру даних.

Ідея мережі Кохонена належить фінському вченому Тойво Кохонену. Принцип роботи цих мереж полягає у введенні в правило навчання нейрона інформації про його розміщення, тобто складаються карти розміщення нейронів.

Самоорганізаційні карти Кохонена використовуються для моделювання, прогнозування, пошуку закономірностей у великих масивах даних, виявлення наборів незалежних ознак і стиснення інформації.

2. Аналіз даних (data mining), заснований на нечіткій нейронній мережі. В основі нечітких нейронних мереж лежить ідея використання наявної вибірки даних для визначення параметрів функцій приналежності, висновки формулюються на основі апарату нечіткої логіки, а для знаходження параметрів функцій приналежності використовуються алгоритми навчання нейронних мереж. Такі системи можуть використовувати заздалегідь відому інформацію, навчатися, здобувати нові знання, прогнозувати часові ряди, виконувати класифікацію образів. Але однією з головних переваг є наочність роботи такої мережі для користувача.

Кожен із розглянутих типів нейромереж має свої переваги і недоліки щодо інтелектуального аналізу даних, тож доцільно порівняти нейромережу Кохонена у групі типів інтелектуального аналізу даних, заснованих на нейронних мережах.

Порівняння продемонстроване в таблиці 2.

Із таблиці видно, що і мережі Кохонена, і нечіткі нейронні мережі мають переваги і недоліки.

Основна відмінність мереж Кохонена від інших типів нейронних мереж полягає в наочності і зручності використання. Ці мережі дають змогу спростити багатовимірну структуру, їх можна вважати одним із методів проектування багатовимірного простору у простір із більш низькою розмірністю. Інша принципова відмінність мереж Кохонена від інших

Таблиця 1

	BIRCH	**CURE**	**CHAMELEON**	**ROCK**	**KOHONEN**
Великі обсяги даних	+	+	+	+	+
Стійкість до шуму	+	+–	+–	–	+
Масштабованість	+	–	+	+	+
Визначення кількості кластерів	+	–	+	-	+–
Кластери довільного розміру та щільності	–	+–	+	–	+–

Таблиця 2

Тип нейронної мережі	**Область використання**	**Переваги**	**Недоліки**
Мережа Кохонена	Класифікація, кластерний аналіз, прогнозування, стиснення даних	Стійкість до зашумлених даних, некероване навчання, швидке навчання, можливість візуалізації, можливість спрощення багатовимірної структури	Евристичність алгоритму навчання
Нечітка нейронна мережа	Класифікація, прогнозування	Хороша сумісність, швидке навчання, інтерпретованість накопичених знань, наочність роботи, легкість визначення розміру мережі, допустимість зашумлених і неточних даних, здатність апроксимувати функції будь-якого ступеня нелінійності, паралельні обчислення	Апріорне визначення компонентів

моделей нейронних мереж – некероване або неконтрольоване навчання, що дає змогу задавати лише значення вхідних змінних.

Висновки з проведеного дослідження. Підсумовуючи викладений матеріал і міркування, можна дійти висновку про те, що самоорганізуюча нейронна мережа Кохонена може бути однією з основ адекватного алгоритму порівняно з іншими типами нейромереж, призначеними для кластерного аналізу даних.

Мережі Кохонена принципово відрізняються від всіх інших типів мереж. Тоді як всі інші мережі призначені для завдань із керованим навчанням, мережа Кохонена головним чином розрахована на некероване навчання, а це означає, що мережа вчиться розуміти саму структуру даних.

Отже, самоорганізуюча нейронна мережа сьогодні є сильним інструментом у сфері алгоритмів для кластерного аналізу даних та конкурує із сучасними алгоритмами, але жодна з наявних чистих моделей не відповідає сучасним вимогам.

У подальшому автор не відкидає бажання звернути увагу на такі алгоритми, як:

1) графо-орієнтований алгоритм CHAMELEON, що виключає проблему невизначеності кількості кластерів;

2) алгоритм К-середніх – метод жорсткої кластеризації. Це означає, що точка даних може належати тільки одному кластеру і для приналежності кожної точки даних цього кластера обчислюється одне значення ймовірності.

3) максимізація очікувань (ЕМ) – це метод м'якої кластеризації. Це означає, що точка даних завжди належить до кількох кластерів і для всіх можливих поєднань точок даних із кластерами обчислюються ймовірності.

Варто зазначити, зокрема, що моделі кластеризації «Майкрософт» використовують алгоритм масштабованої максимізації очікування. Цей алгоритм використовується за замовчуванням, оскільки він має декілька переваг порівняно з методом кластеризації К-середніх:

– не вимагає більше одного перегляду бази даних;

– працює навіть за обмеженого обсягу оперативної пам'яті,

– може використовувати однопрохідний курсор;

– за продуктивностю випереджає методи, що вимагають вибірки.

Одним із найсуттєвіших недоліків нейромережі Кохонена є те, що відповідний алгоритм не передбачає визначення кількості кластерів. Але він здатен функціонувати в умовах перешкод завдяки тому, що число кластерів фіксовано завчасно.

ЛІТЕРАТУРА:

1. Дюк В.А. Data Mining – интеллектуальный анализ данных // Информационные технологии: сайт. – URL: http://www.inftech.webservis.ru/it/database/datamining/ar2.html

2. Xianjun Ni Research of Data Mining Based on Neural Networks // World Academy of Science, Engineering and Technology. – 2008. – № 39. – P. 381-384.

3. Xu R. and Wunsch D. II. Survey of Clustering Algorithms. IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005, pp. 645ñ678.

4. Мандель И.Д. Кластерный анализ. М.: Финансы и статистика, 1988. ñ 176 с.

5. Стариков А. Практическое применение нейронных сетей для задач классификации (кластеризации), http://www.basegroup.ru/neural/practice.htm, январь 2000.

6. George Karypis. Chameleon: Hierarchical Clustering Using Dynamic Modeling / George Karypis, Eui-Hong (Sam) Han, Vipin Kumar // Computer. – 1999. – Vol. 32, N 8. – P. 68-75.

7. Ярушкина Н.Г. Основы теории нечетких и гибридных систем: учеб. пособие. – М.: Финансы и статистика, 2004. – 320 с.