УДК 004.032.26:551.5

В.В. Казимир, д-р техн. наук
О.Ю. Герасименко, аспірант
Чернігівський національний технологічний університет, м. Чернігів, Україна

ЗАСТОСУВАННЯ МЕТОДІВ DATA MINING ДЛЯ АНАЛІЗУ МЕТЕОРОЛОГІЧНИХ ПОКАЗНИКІВ

В.В. Казимир, д-р техн. наук
О.Ю. Герасименко, аспирант
Черниговский национальный технологический университет, г. Чернигов, Украина

ПРИМЕНЕНИЕ МЕТОДОВ DATA MINING ДЛЯ АНАЛИЗА МЕТЕОРОЛОГИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Volodymyr Kazymyr, Doctor of Technical Sciences
Oksana Herasymenko, PhD student
Chernihiv National University of Technology, Chernihiv, Ukraine

USING DATA MINING TECHNIQUES FOR METEOROLOGICAL FIGURES ANALYSIS

Розглянуто підходи у використанні методів Data Mining (DM) для оброблення та аналізу метеорологічних даних. Коротко описано технологію Knowledge Discovery in Databases (KDD) та представлено опис методів DM як одного з етапів KDD. Наведено приклад короткострокового прогнозування температури повітря для м. Чернігова з використанням нейромереж (вхідними даними були дані метеостанції м. Чернігова за 2013–2015 роки).

Ключові слова: *Knowledge Discovery in Databases, Data Mining, нейронні мережі, багатошаровий персептрон, прогноз метеорологічних показників.*

Рассмотрены подходы использования методов Data Mining (DM) для обработки и анализа метеорологических данных. Кратко описано технологию Knowledge Discovery in Databases (KDD) и выполнено описание методов Data Mining как одного из этапов KDD. Приведен пример краткосрочного прогноза температуры воздуха для г. Чернигова с использованием нейросетей (входящими данными были данные метеостанции г. Чернигова за 2013–2015 годы).

Ключевые слова: *Knowledge Discovery in Databases, Data Mining, нейронные сети, многослойный персептрон, прогноз метеорологических показателей.*

This article describes several approaches to using Data Mining techniques for processing and analysis of the meteorological data. The Process of Knowledge Discovery in Databases (KDD) and some method of Data Mining (DM), one of the KDD's step, are briefly described in the article. As an example of using Data Mining techniques for meteorological figures analysis a temperature forecasting by neural networks is performed in this article (as input data we used meteorological figures from Chernihiv weather station for 2013-2015 years).

Key words: *Knowledge Discovery in Databases, Data Mining, neural networks, multilayer perceptron, meteorological figures forecasting.*

Постановка проблеми. Об'єм накопичених даних метеорологічних спостережень з кожним роком збільшується. Це створює умови для формування метеопрогнозів, ґрунтуючись тільки на цих даних, наприклад, за допомогою методів Data Mining (DM). Отримані в результаті аналізу метеоданих знання можуть бути корисними як для розуміння тенденцій зміни клімату, так і для тих сфер діяльності людини, які залежать від погодних умов, включаючи короткострокові прогнози. Використання методів Data Mining дозволяє виявити певні закономірності, характерні для тієї чи іншої території або проміжку часу. Наприклад, у [3] застосовується алгоритм DBSCAN для визначення регіонів Туреччини, подібних за своїми температурними характеристиками.

Мета статті. Метою статті є дослідження можливостей технології Knowledge Discovery in Databases (KDD) та методів DM, як одного з етапів KDD, для формування короткострокових прогнозів на основі аналізу метеопоказників за попередні тривалі періоди.

Аналіз досліджень і публікацій. Оброблення та аналіз метеорологічних даних з використанням Data Mining здійснювалися багатьма дослідниками. До основних задач, які вирішувалися із застосуванням DM, відносяться: прогнозування метеорологічних

показників; виявлення закономірностей, характерних для певної території та/або проміжку часу, прогнозування екстремальних погодних явищ.

Це продемонстровано в деяких роботах закордонних авторів. Так, у [5] розглядається побудова та аналізується використання нейромережі для погодинного прогнозу швидкості вітру в м. Фару (Португалія). Для аналізу було взято дані зі швидкості вітру за два роки. За результатами досліджень найкращі результати прогнозування продемонстрував багатошаровий персептрон (три шари по 14-15-1 нейронів відповідно) з прямим зв'язком. Вхідними даними для нейромережі були лише значення швидкості вітру.

У [9] представлено прогнозування температури повітря також із використанням нейромереж, причому дані було розбито по сезонах і для кожного сезону будувалась окрема нейромережа. Вхідними даними були не лише значення температури повітря, але й певні інші показники.

Прогнозування максимальної, мінімальної та середньої добових температур повітря на рік з використанням різних методів DM виконано у [10]. Застосовувалися такі алгоритми, як класична лінійна регресія, M5, M5rules, IB3, а також адитивна регресія та нейронні мережі. Крім того, у цій роботі показано, що для отримання прогнозу задовільної якості достатньо даних за два попередні роки.

У [12] використано нейронні мережі різної конфігурації та два ансамблі нейромереж для погодинного прогнозу на добу температури повітря, швидкості вітру і вологості повітря. Найкращий результат був отриманий із застосуванням рекурентної нейронної мережі Елмана, а серед ансамблів нейронних мереж кращий прогноз показав ансамбль, сформований за принципом «winner-take-all».

Для прогнозування екстремальних погодних явищ у дослідженнях [2; 12; 18] використовувалися різноманітні методи DM, оскільки прогнозування різних явищ вимагає неоднакових вхідних даних та підходів до проведення аналізу.

Технологія Knowledge Discovery in Databases (KDD). Ця технологія почала активно розвиватись на початку 1990-х років, що було зумовлено такими факторами:

– збільшення місткості накопичувачів даних та зменшення їх вартості, що сприяло стрімкому зростанню об'ємів накопичених «сирих» даних. Під «сирими» даними розуміють різнорідні та необроблені дані;

– прогрес у сфері створення комп'ютерної техніки, що сприяло підвищенню обчислювальної потужності техніки і загальній комп'ютеризації виробництва та бізнес-процесів;

– стрімкий розвиток інформаційно-комунікаційних технологій;

– впровадження та широке використання Інтернету в усіх сферах людської діяльності.

KDD може проводитись за двома напрямами: перевірка гіпотези користувача (verification) або видобуток нових знань (discovery), який, у свою чергу, може мати на меті прогнозування (prediction) або опис (description) об'єкта дослідження.

KDD – це нетривіальний процес виявлення значущих, раніше невідомих, потенційно корисних і в кінцевому результаті зрозумілих паттернів у даних; під даними слід розуміти множину фактів, а паттерн – це описаний певною мовою вираз для позначення підмножини даних або моделі, яка може бути використана до підмножини даних [4]. KDD дозволяє виявити в наборі фактів нові корисні багатоаспектні залежності між даними. Цей процес є ітеративним та інтерактивним і складається з деяких етапів (рис. 1).

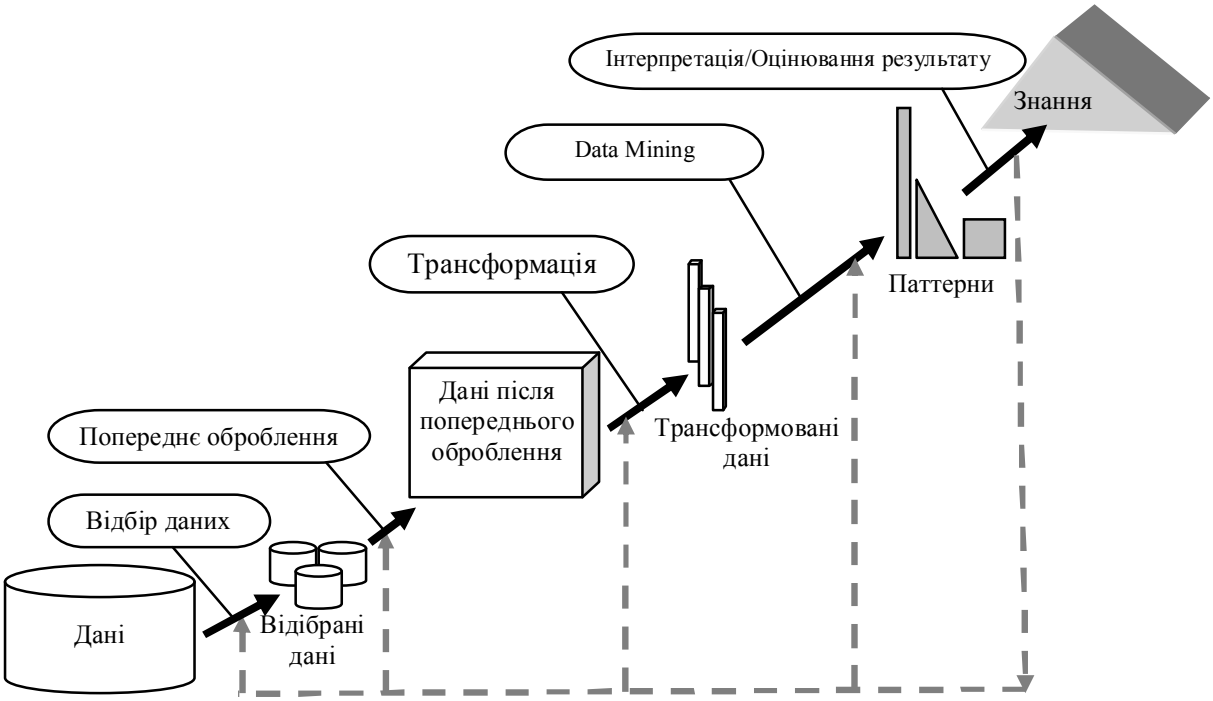

Рис. 1. Схематичне зображення процесу KDD [4]

Результатом процесу KDD є знання, а DM - це один з етапів процесу KDD, який передбачає використання алгоритмів аналізу даних та виявлення інформації, з прийнятними обмеженнями обчислювальної ефективності, для синтезу специфічного переліку паттернів (або моделей) із даних [4]. Іншими словами, DM – це сукупність алгоритмів і методів, призначених для виявлення та представлення у зрозумілому для аналітика вигляді відношень між даними чи закономірностей у даних, але поряд з цим не проводиться оцінювання значущості та корисності отриманих результатів. Таким чином, набір паттернів, отриманий у результаті етапу DM, не є результатом процесу KDD загалом, оскільки отриманню знань передує інтерпретація паттернів, оцінювання їх якості та придатності.

Data Mining у своєму арсеналі має значну кількість методів і алгоритмів, оскільки ця технологія виникла на стику кількох галузей науки, зокрема штучного інтелекту, машинного навчання, розпізнавання образів, статистики, бази даних, візуалізації даних, високопродуктивного обчислення та ін. Для зручності алгоритми та методи DM краще розглядати залежно від задач, для вирішення яких вони застосовуються. Перелік задач Data Mining у різних джерелах може дещо різнитися, тому розглянемо ті задачі, які більшість дослідників вважають основними. До них належать:

– класифікація (classification) та регресійний аналіз (regression);
– кластеризація (clustering);
– пошук частих моделей (frequent pattern mining).

Класифікація та регресія. Під поняттям «класифікація», як правило, мають на увазі розподіл елементів даних в один з декількох наперед визначених класів елементів. У [1] задача класифікації представлена більш узагальнено й аналогічно задачі регресії, тобто як така, метою якої є «визначення значення залежної змінної деякого об’єкта на основі значень інших змінних, які характеризують цей об’єкт» [1, с. 102]. З метою класифікації можуть бути використані неоднакові алгоритми, що використовують різні способи вирішення задачі. У [20] розглядаються такі підходи класифікації: імовірнісна класифікація (probabilistic classification), класифікація з використанням дерев рішень (decision

tree classifier), лінійний дискримінантний аналіз (linear discriminant analysis), метод опорних векторів (support vector machines). Результати класифікації (чи регресії) можуть бути представлені у вигляді класифікаційних правил, дерев рішень, математичних функцій [1].

Класифікаційні правила складаються з двох частин – умови та висновку. В умові виконується перевірка значень однієї чи декількох незалежних змінних, а у висновку, відповідно, вказується значення залежної (прогнозної) змінної або розподіл її імовірності за класами. До переваг класифікаційних правил можна віднести простоту сприйняття людиною, адже вони можуть бути записані на природній для людини мові, а також їх відносну незалежність – можливість додавати у набір правил нові правила без зміни вже наявних. Суттєвим недоліком класифікаційних правил вважається можлива суперечливість їх одне одному, коли характеристики якогось об’єкта задовольняють умови кількох правил із різними висновками. Для побудови класифікаційних правил можуть бути використані алгоритми 1R-алгоритм [1], Naive Bayes [1; 6; 11; 14; 20] та інші.

Представлення класифікаційних правил у вигляді деревовидної послідовної структури є деревом рішень. Кожен вузол такого дерева містить перевірку деякої незалежної змінної (можуть також порівнюватись незалежні змінні між собою, обчислюватись певна функція від однієї чи кількох незалежних змінних), а кожен лист дерева є значенням залежної змінної, тобто класом. Пройшовши від кореня дерева до листа через вузли з умовами, які задовольняють незалежні змінні об’єкта класифікації, визначається клас, до якого належить об’єкт. Важливо зазначити, що маючи дерево рішень, його можна записати у вигляді класифікаційних правил, але зворотне перетворення можливе не завжди. Існує багато алгоритмів побудови дерев рішень, зокрема ID3 [1; 14; 16], C4.5 [1; 11; 15], CART [11], CHAID [17], SPRINT [14] та інші. Основними характеристиками алгоритмів побудови дерев рішень є вид розщеплення (бінарне чи множинне), критерій розщеплення, процедура скорочення гілок дерева, можливість оброблення пропущених значень. Вибір того чи іншого алгоритму залежить від багатьох факторів, тому жоден з алгоритмів не вважається найкращим.

До імовірнісної класифікації у [20] віднесено Байєсівську класифікацію та класифікацію методом k-найближчих сусідів. Байєсівська класифікація (повна) [20] використовує формулу Байєса для визначення класу, до якого належить об’єкт, як класу з максимальною апостеріорною вірогідністю. Обчислити апостеріорну імовірність, згідно з формулою Байєса, можна за апріорними ймовірностями та умовними за класом щільності. Проста Байєсівська класифікація (Naive Bayes classifier) [1; 6; 11; 14; 20] ґрунтується на припущенні, що всі змінні статистично незалежні між собою, і покликана спростити обчислення в порівнянні з повною Байєсівською класифікацією.

Класифікація методом k-найближчих сусідів (k-nearest neighbors classifier, KNN classifier) [6; 7; 14; 16; 20] полягає у визначенні класу, до якого належить найбільше найближчих сусідів об’єкта, клас якого визначається, і вважається, що новий об’єкт також належить до цього класу. Найближчі сусіди визначаються, переважно, за допомогою обчислення відстані між точками у просторі за формулою Евкліда, хоча також можливі інші метрики. Число k позначає кількість найближчих сусідів, які використовуються для класифікації.

Лінійний дискримінантний аналіз [20] використовується для пошуку лінійної комбінації ознак, яка найкращим чином розділяє об’єкти на класи, або, іншими словами, полягає у знаходженні такого вектора, після проекції на який об’єкти можна розподілити між класами з максимальною сепарацією (максимальним розділенням). Як критерій розділення під час аналізу використовується лінійний дискримінант Фішера. Також лінійний дискримінантний аналіз може виконуватись на базі функцій-ядер [20].

Метод опорних векторів (SVM) [1; 6; 11; 14; 20] ґрунтується на ідеї, що найкращим способом розмежування точок в m-мірному просторі є m-1 гіперплощина, рівновіддалена від точок, які належать різним класам [1]. Основним завданням методу є пошук серед усіх можливих площин такої, яка рівновіддалена від крайніх об'єктів кожного з класів, і такі об'єкти називаються опорними векторами. Найкращим варіантом використання SVM класифікатора є той, коли простір між межами класів пустий. Проте також розроблені принципи застосування методу опорних векторів для класів, які перекриваються [20]. У разі неможливості лінійного розділення об'єктів класів може бути використаний метод опорних векторів на базі функцій-ядер [1; 20]. Перевагами методу вважаються [1]: теоретична і практична обґрунтованість; узагальнений підхід до розв'язку багатьох задач завдяки застосуванню різних функцій-ядер; сталі розв'язки, незалежні від локальних мінімумів; відсутня проблема «перенавчання» (overfitting); може використовуватись при будь-яких розмірностях. Недоліками методу опорних векторів є [1]: порівняно невисока продуктивність; відсутність рекомендацій щодо вибору параметрів та вибору ядра; побічні ефекти нелінійних перетворень; труднощі в інтерпретації результату; застосування для невеликої кількості векторів.

Також для класифікації можуть бути використані нейронні мережі, генетичні алгоритми, теорія наближених множин, теорія нечітких множин та інші.

Оцінити точність класифікації можна за допомогою так званої крос-перевірки (cross-validation) – процедури перевірки точності класифікації на даних із тестової множини. Якщо точність класифікації тестової множини приблизно однакова з точністю класифікації навчальної вибірки, то вважається, що отримана модель пройшла крос-перевірку. В [6] розглядаються способи покращення точності класифікації, а саме ансамблевий метод.

Кластеризація. На відміну від класифікації, де кількість класів задана наперед, кластеризація (кластерний аналіз) полягає в розподілі заданого набору об'єктів на групи схожих між собою об'єктів, які називаються кластерами. Кількість кластерів заздалегідь невідома, тому завдання кластерного аналізу – визначення кількості кластерів і віднесення кожного з об'єктів набору даних до одного (або декількох) кластера (кластерів). Отже, модель, отримана в результаті кластеризації, повинна описувати як безпосередньо кластери, так і належність об'єкта даних до одного (чи декількох) з цих кластерів.

Методи кластеризації поділяються на неієрархічні та ієрархічні, останні у свою чергу поділяються на агломеративні та дивізимні [1], у кожній з цих груп є безліч алгоритмів та підходів. Ієрархічні методи представляють результат у вигляді дерева, вкладених один в одного кластерів, що дає змогу найбільш повно представити структуру кластерів [1]. При ієрархічній кластеризації немає необхідності визначати кількість кластерів, тоді як неієрархічні методи в основному спрямовані на добір числа кластерів та визначення у процесі цього добору їх оптимальної кількості. Використовуючи різні методи кластеризації для тих самих даних, можна отримати різні результати, що вважається прийнятним. Під час вибору алгоритму кластеризації важливо знати сильні та слабкі сторони алгоритму, а також враховувати природу даних, які будуть оброблятися.

Розглянемо деякі алгоритми кластерного аналізу. Неієрархічна кластеризація ґрунтується на мінімізації певної цільової функції, яка визначає оптимальне розподілення об'єктів за кластерами. Класичним алгоритмом неієрархічної кластеризації є алгоритм k-середніх (k-means). Повний опис цього алгоритму можна знайти в [8], також він описується в [1; 6; 7; 11; 20]. Метою алгоритму є побудувати кластери таким чином, щоб їх середні якомога більше відрізнялись між собою. Кількість кластерів k задається до початку аналізу, тому насамперед необхідно мати гіпотезу про їх кількість. Якщо значення k неможливо спрогнозувати, то спочатку береться значення 2, а потім 3, 4, 5 і т. ін. З

набору даних довільним чином вибираються k точок, які вважаються центрами кластерів. Потім усі об'єкти розбиваються на k груп, кожна з яких концентрується навколо одного з центрів. Далі обчислюються нові центри кластерів і знову проводиться формування k груп. Цей процес повторюється до того часу, доки центри кластерів не перестануть змінюватись. Алгоритм k-середніх став базовим для інших неієрархічних алгоритмів кластеризації, наприклад, для алгоритму PAM (partitioning around medoids) [14]. Алгоритм k-середніх на базі функцій-ядер розглядається в [20]. У результаті узагальнення алгоритму k-середніх розроблено алгоритм Fuzzy C-Means, в якому кластери вважаються нечіткими множинами і належність кожного об'єкта набору даних до того чи іншого кластера визначається за критерієм максимуму належності до цього кластера [1]. Алгоритм Fuzzy C-Means шукає кластери сферичної форми, на його основі було розроблено алгоритм кластеризації за Гюстафсоном-Кесселем, який шукає кластери у формі еліпсоїдів. Як зазначено у [1], недоліком названих вище алгоритмів є те, що вони шукають кластери певної форми та ґрунтуються на аналізі відстані від точки до центра кластера, а не на аналізі взаємного розміщення точок.

Також серед методів неієрархічної кластеризації необхідно відзначити EM-алгоритм (Expectation-Maximization) [6; 11; 20], який припускає наявність для кожного кластера функції щільності розподілу імовірності з відповідними значеннями математичного очікування та дисперсії і завданням алгоритму є знайти параметри розподілів за принципом максимуму правдоподібності. Однією з найважливіших переваг цього методу є можливість масштабування (він дозволяє проводити оброблення великих масивів даних).

Крім того, до неієрархічних методів кластеризації належать методи, які ґрунтуються на аналізі концентрації (щільності розподілу) об'єктів (density-based clustering). Основний принцип такого підходу – об'єкти кластера, розміщені з певною щільністю, яка більша, ніж щільність об'єктів за межами кластера. Таку класифікацію реалізують алгоритми DBSCAN [6; 14; 20], DENCLUE [6; 14; 20], OPTICS [6; 14].

Ієрархічна кластеризація будує так звані дендрограми, тобто деревовидні структури з відображенням вкладеності кластерів. Під час побудови використовуються агломеративні методи (AGNES, Agglomerative Nesting) – виконують послідовне об'єднання менших кластерів у більші - або дивізимні методи (DIANA, Divisive Analysis) – розділяють більші кластери на менші. До перших можна віднести алгоритми CURE [14], ROCK [14], CAMELEON [6] та інші, а до других – BIRCH [6; 14], хоча слід зауважити, що в цих алгоритмах ієрархічні методи інтегровані з іншими. З основними принципами агломеративних та дивізимних методів можна ознайомитись у [6; 7; 14; 20].

Окремо слід зупинитися на оцінюванні результатів кластерного аналізу. В [1] розглядається поняття «якість кластеризації», на основі якої ґрунтується вибір оптимального рішення у процесі кластерного аналізу. Якість кластеризації – ступінь наближення результату кластеризації до ідеального рішення [1]. Для оцінювання якості кластеризації використовуються формальні параметри, оцінювання може проводитись за різними показниками. Основні показники, за якими може проводитись оцінювання якості кластеризації:

– показники чіткості. Вони набувають максимального значення при найбільш чіткому розділенні об'єктів даних. До них належать коефіцієнт розбиття, модифікований коефіцієнт розбиття, індекс чіткості;

– ентропійні критерії. Кластеризація вважається якіснішою тоді, коли значення ентропії найнижче. Це буде тоді, коли ступінь належності елементу даних до одного кластера найбільша, а до інших - найменша. До цих критеріїв належать ентропія розбиття, модифікована ентропія;

– інші показники, наприклад, показник компактності та ізольованості, індекс ефективності і т. ін.

Для більш детального ознайомлення з критеріями оцінювання результатів кластерного аналізу можна звернутись до [20].

Використовуючи критерії оцінювання якості кластеризації, в алгоритмі кластерного аналізу можна закласти певний адаптивний механізм вибору оптимального розв'язку серед усіх можливих, що приводить до поняття адаптивної кластеризації. Це найбільш характерно для неієрархічних методів кластеризації, де важко спрогнозувати результуючу кількість кластерів.

Також для кластеризації застосовуються такі підходи, як графові алгоритми кластеризації, нейронні мережі, генетичні алгоритми, еволюційні алгоритми, спектральна кластеризація, кластеризація на базі дерев рішень та інші. На сьогодні важливою вимогою до алгоритмів кластеризації вважається можливість їх масштабування, тобто здатність алгоритмів обробляти надвеликі об'єми даних. Прикладами алгоритмів, здатних до масштабування (scalable), є CLARA [6], CLARANS [6], CLOPE [19], CURE, BIRCH та багато інших.

Пошук частих моделей. Частими паттернами можна назвати певні елементи чи групи елементів деякої предметної області, які часто повторюються, а часті моделі призначені для опису частих паттернів.

Аналіз частих наборів (frequent itemsets mining) ґрунтується на понятті транзакції - наборі даних із бази даних, тобто існує деяка множина елементів, із яких сформовано набори даних бази даних і ці набори є об'єктами аналізу. Найпростішим прикладом таких наборів є перелік товарів у чеку. Аналіз частих наборів для виявлення якихось закономірностей називають пошуком асоціативних правил. Основною метою побудови асоціативних правил є виявлення наборів даних, які часто повторюються серед великої кількості таких наборів. У результаті будуються так звані асоціативні правила виду

якщо (умова), то (результат), де

умова – набір об'єктів, з якими асоціюються об'єкти, які містяться в *результаті* правила.

Побудовані правила не обов'язково несуть корисну інформацію, тому застосовуються спеціальні величини для оцінювання їх корисності. Згідно з [1] до них належать:

– підтримка (support) – демонструє відсоток транзакцій, які підтримують це правило;

– достовірність (confidence) – відображає імовірність того, що наявність у транзакції одного набору об'єктів приводить до наявності іншого набору даних;

– покращення (improvement) – показує, чи корисніше застосування правила від простого відгадування.

Для пошуку асоціативних правил застосовується алгоритм Apriori [1; 11; 14; 20] та його різновиди AprioriTid [1; 6], MSAP [1], а також алгоритми Eclat [20], dEclat [20], FPGrowth [6; 20].

Опрацювання частих наборів великої кількості елементів пов'язане зі зростанням обчислювальної складності задачі. У [20] розглядається один з підходів до вирішення цієї проблеми, який полягає у формуванні так званого згущеного представлення частих наборів, яке узагальнює їх значущі характеристики.

Аналіз частих послідовностей (frequent sequences mining) має на меті виявлення залежностей між пов'язаними подіями у базі даних, тобто є певні послідовності подій і необхідно виявити закономірності у цих послідовностях. Отже, на відміну від частих наборів, у цьому випадку визначена певна послідовність елементів у наборі, критерієм упорядкування послідовності може бути, наприклад, час. Аналіз частих послідовностей допомагає виявити тенденції за часом або позицією розміщення елементу в наборі даних [20]. Для аналізу частих послідовностей існують відповідні алгоритми, наприклад, Spade [20], PrefixSpan [20] та інші.

У зв'язку зі значним поширенням різноманітних мереж у сучасному світі виникає все більше необхідності в аналізі даних, представлених у вигляді графів, наприклад, у вигляді графів зручно представляти блоги, соціальні мережі, Інтернет, біологічні системи та багато іншого. Метою такого аналізу є виявлення цікавих та корисних підграфів у одному великому графі або із бази даних графів. Основні принципи аналізу графів описано в [20].

Як і під час кластеризації, знайдені в результаті аналізу правила та часті моделі необхідно оцінити на предмет значущості. Вище було розглянуто деякі показники оцінювання асоціативних правил, більш детально показники оцінювання правил та частих моделей розглядаються в [20].

Крім названих вище задач DM, можна зазначити ще такі:

– виявлення змін і відхилень (change and deviation detection), метою якого є виявлення у наборі даних паттернів, не характерних для цього набору даних, або виявлення важливих змін у порівнянні з попереднім заміром параметрів;

– підведення підсумків (summarization) дозволяє знайти компактний опис підмножини даних із загального набору даних. Також часто використовується для реферування текстів;

– візуалізація або візуальний аналіз даних (visualization, visual mining), метою якого є зображення даних у певній візуальній формі, завдяки чому аналітик може зрозуміти їх суть і зробити певні висновки.

Таким чином, DM у своєму арсеналі має значну кількість методів, у тому числі нейронні мережі, генетичні алгоритми, еволюційні алгоритми, а вибір того чи іншого методу (або комплексу методів) залежить від задачі, яку необхідно вирішити.

Аналіз метеорологічних даних засобами Data Mining. Розглянемо прогнозування температури повітря з використанням нейромереж. Для дослідження виберемо дані метеорологічної станції, розміщеної в м. Чернігів (WMO_ID=33135). Дані являють собою заміри через кожні три години різних метеорологічних показників, зокрема температура повітря, атмосферний тиск, відносна вологість повітря, напрям вітру, швидкість вітру та багато інших. Для аналізу було вибрано дані температури повітря за період з 01.01.2013 по 13.10.2015, загалом 8125 значень.

Підготовка даних. Слід зазначити, що для метеорологічних даних характерна низька якість даних, тому перш ніж почати аналіз, їх необхідно перевірити на якість та усунути виявлені недоліки, адже від цього може значно залежати отриманий результат. Тому спершу було оцінено якість вхідних даних. Викидів та екстремальних значень виявлено не було, пропусків виявлено 35, тобто 0,43 %. Оскільки метеорологічні дані являють собою часові ряди, то для заповнення пропусків вибрано метод інтерполювання.

Прогнозування температури здійснювалося з використанням нейронної мережі, а саме багатошарового персептрону. В середовищі STATISTICA було вибрано стратегію «Автоматизована нейронна мережа» та побудовано 250 різних мереж і визначено 5 із найкращими характеристиками (у цьому випадку застосовувалась функція помилки «Сума квадратів»). Вхідними даними для нейромережі були 4 послідовні значення температури повітря; кількість нейронів проміжного шару варіювалася від 3 до 25; також змінювалася функція активації нейронів проміжного та вихідного шарів (доступними були функції: лінійна, логістична, гіперболічний тангенс, експонента); кількість вихідних значень – одна – температура повітря. Набір даних для навчання мережі становив 80 %, тестовий набір – 20 %. У результаті було вибрано 5 мереж, характеристики яких представлено на рис. 2.

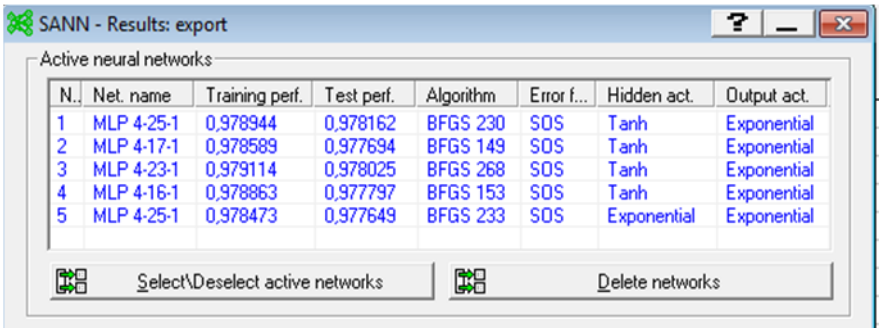

Рис. 2. Параметри нейронних мереж, відібраних для прогнозування температури повітря

Відібрані нейронні мережі були використані для прогнозування температури повітря на 15 кроків вперед. Прогноз робився, починаючи з 8110 кроку, щоб отримані результати можна було порівняти з фактичним значенням температури повітря. На рис. 3 представлено графіки фактичної температури повітря T та прогнозної температури повітря за даними кожної із нейронних мереж.

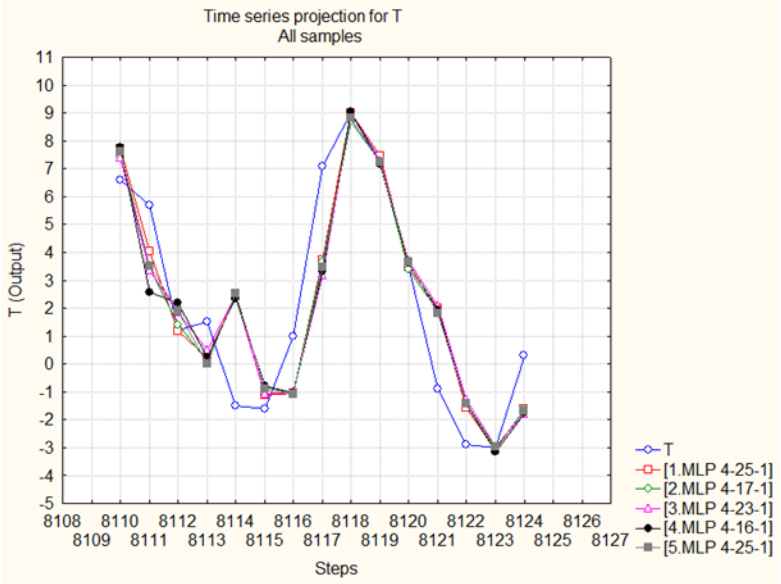

Рис. 3. Графік фактичної та прогнозних температур повітря

Згідно з рис. 3 можна сказати, що відібрані нейромережі дають дуже близькі прогнози температури повітря. У таблиці представлено коефіцієнти кореляції та значення середньої квадратичної похибки для кожного із прогнозів.

Таблиця

Коефіцієнти кореляції та середня квадратична похибка для прогнозів кожної з мереж

Нейромережа	Середня квадратична похибка (°C)	Коефіцієнт кореляції
1.MLP-4-25-1	1,838866	0,882303914
2.MLP-4-17-1	1,899579	0,873162009
3.MLP-4-23-1	1,980348	0,861381256
4.MLP-4-16-1	2,035408	0,854014153
5.MLP-4-25-1	1,941979	0,867174574

Аналогічним чином можна виконати прогноз й інших метеорологічних параметрів, таких як атмосферний тиск, відносна вологість повітря, швидкість вітру.

Висновки і пропозиції. У статті на прикладі прогнозування температури повітря із застосуванням нейромережі продемонстровано використання Data Mining для прогнозу метеорологічних даних, а саме виконано короткостроковий прогноз температури повітря на 45 годин з інтервалом у 3 години. Найбільша абсолютна похибка прогнозування

становила 4,04 °С (нейромережа MLP-4-25-1), середня абсолютна похибка становить 1,47 °С, що підтверджує ефективність запропонованого прогнозу.

Список використаних джерел

1. *Анализ* данных и процессов : учеб. пособие / [А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров]. – [3-е изд., перераб. и доп.]. – СПб. : БХВ-Петербург, 2009. – 512 с.

2. *Bartok J.* Data Mining for fog prediction and low clouds detection / J. Bartok, F. Babiˇc, P. Bedn´ar, J. Paraliˇc, J. Kov´aˇc, I. Bartokov´a, L. Hluch´y, M. Gera // Computing and Informatics. – 2011. - Vol. 30. – P. 1441–1464.

3. *Bilgin T. A.* data mining application on air temperature database / T. A. Bilgin, A. Çamurcu // Advances in Information Systems. - Springer Berlin Heidelberg. – 2005. - P. 68–76.

4. *Fayyad U.* From Data Mining to Knowledge Discovery in Databases / U. Fayyad, G. Piatetsky-Shapiro, P. Smyth // Al Magazine. – 1996. – № 17 (3): FALL. - P. 37–54.

5. *Fonte P. M.* Wind speed prediction using artificial neural networks / P. M. Fonte, G. X. Silva, J. C. Quadrado // Proceedings of the 6th WSEAS Int. Conf. on neural networks, Lisbon, Portugal, June 16–18, 2005. – P. 134–139.

6. *Han J.* Data Mining: concepts and techniques / J. Han, M. Kamber, J. Pei. – 3rd ed. – Elsevier, 2011. – 744 p.

7. *Hand D.* Principles of Data Mining / D. Hand, H. Mannila, P. Smyth. – Cambridge, Massachusetts : MIT Press, 2001. – 546 p.

8. *Hartigan J. A.* A K-Means clustering algorithm / J. A. Hartigan, M. A. Wong // Journal of the Royal Statistical Society. Series C (Applied Statistics). – 1979. - Vol. 28, No.1. - P. 100–108.

9. *Hayati M.* Temperature forecasting based on neural network approach / M. Hayati, Z. Mohebi // World Applied Sciences Journal. – 2007. – Vol. 2, Num. 6. – P. 613–620.

10. *Kotsiantis S.* Using data mining techniques for estimating minimum, maximum and average daily temperature values / S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias // International Journal of Mathematical, Physical and Engineering Sciences. – 2008. - Vol. 1, Num. 1. - P. 16–20.

11. *Kumar V.* The Top ten algorithms in Data Mining / V. Kumar, X. Wu. – Taylor&Francis Group, LLC, 2009. – 2008 p.

12. *Li X.* Real-time storm detection and weather forecast activation through data mining and events processing / X. Li, B. Plale, N. Vijayakumar, R. Ramachandran, S. Graves, H. Conover // Earth Sci Inform. – 2008. – Vol. 1. – P. 49–57.

13. *Maqsood I.* An ensemble of neural networks for weather forecasting / I. Maqsood, M. R. Khan, A. Abraham // Neural Computing & Applications. 2004. – Vol. 1, Num. 2. – P. 112–122.

14. *Mitra S.* Data mining: multimedia, soft computing and bioinformatics / S. Mitra, T. Acharya. – John Wiley&Sons, Inc., 2003. – 424 p.

15. *Quinlan J. R.* C4.5: Programs for machine learning. - Morgan Kaufmann, Los Altos, 1993. – 303 p.

16. *Quinlan J. R.* Induction of decision trees / J. R. Quinlan // Machine Learning 1, 1986. – P. 81–106.

17. *Rokach L.* Data Mining with decision trees: theory and applications / L. Rokach, O. Maimon. - World Scientific Publishing, 2007. – Vol. 61. – 270 p. – (Series in Machine Perception and Artificial Intelligence).

18. *Shanmuganathan S.* Data Mining methods to generate severe wind gust models / S. Shanmuganathan, Ph. Sallis // Atmosphere. – 2014. – Vol. 5. – P. 60–80.

19. *Yang Y.* CLOPE: A fast and effective clustering algorithm for transactional data / Y. Yang, H. Guan, J. You // Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002. - P. 682–687.

20. *Zaki M. J.* Data mining and analysis: fundamental concepts and algorithms / M. J. Zaki, M. J. Wagner. – NY : Cambridge University Press, 2014. – 593 p.