Підхід щодо отримання нечітких логічних правил із набору статистичних даних

Глоба Л.С.[1],Шелест Є.В. [1], Ляшенко А.В.[1]
([1]КПІ ім. І. Сікорського, E-mail: lgloba@its.kpi.ua, krohaenot522@gmail.com, andrey.lyashenko44@gmail.com)

Швидке збільшення обсягу даних. які передаються мережею і одночасно необхідність їх обробки за умови жорстких вимог до якості обслуговування (швидкості передавання та обробки інформації, достовірності інформації) призвели до появи такого феномену як великі дані, з яким першими й зустрілися телеком оператори [. На сьогоднішній день існує багато рішень щодо обробки великих даних, які застосовують такі глобальні компанії як Google, Amazon,Microsoft та інші та які дозволяють вилучати знання зі значної кількості наборів даних. Таким чином, актуальною залишається проблема зменшення обчислювальної складності задач, націлених на отримання знань із значних обсягів статистичних наборів даних. В даному досліджені пропонується застосування нечітких логічних правил, які отримують автоматизованим шляхом зі статистичних наборів даних, які збирають постійно телеком оператори з метою контролю стану мережі, що дозволяє підвищити достовірність оцінки стану процесів, пов'язаних з передаванням інформації та отриманням доступу до неї у глобальному середовищі, а також зменшити обчислювальну складність під час обробки територіально розподілених даних. В досліджені розглянуто ряд математичних методів побудови нечітких логічних правил, зокрема: очистки, нормалізації, кластеризації та теорія метаграфів. Запропонований підхід щодо побудови нечітких логічних правил дозволяє створити базу нечітких знань, яку можна покласти в основу алгоритмів керування та аналізу складними процесами та системами, зменшуючи обчислювальну складність процесів керування та аналізу технічних систем та врахувати певні нечіткі межі їх граничних станів.

Вели́кі да́ні - це набори інформації настільки великих розмірів, що традиційні способи та підходи аналітики та керування ними не можна застосовувати до них. Індустрія надання телекомунікаційних послуг швидко та динамічно зростає, з'являються нові технології (IoT, M2M, D2D), нові компанії, які застосовують такі технології, впроваджуються нові інформаційно-комунікаційні послуги, що забезпечують автоматизацію виробничих процесів, тощо. Разом з тим зміна мережевої інфраструктури інформаційно-комунікаційних платформ для надання сучасних послуг є досить складною та затратною за часом і грошима задачею. Кількість даних, які генеруються щосекундно становить більше 30 000 Гігабайт [1]. Все це потребує не тільки підтримки та обробки значних обсягів інформації користувачів за допомогою сучасних сервісів обробки значних обсягів даних, але й вилучення саме знань з цих даних та утворення можливостей передбачити майбутні стани досліджуваних процесів або подій.

Процеси аналітичної обробки великих даних характеризуються нечіткими межами, які визначають певні логічні залежності між даними. Такий стан щодо обробки та аналізу великих даних в телеком індустрії вимагає від операторів спеціалізованих сервісів, які б реалізували перехід від обробки даних до отримання знань і від отриманих знань до реалізацій стратегій передбачення.

Таким чином, актуальною є задача удосконалення методів нечіткої логіки для обробки великих даних для побудови системи їх аналізу та обробки із застосуванням мікро-сервісів, що дозволить збільшити масштабованість системи та зменшити завантаженість на сервери, які виконують обробку великих даних.

Для вирішення задач обробки, зберігання та вилучення знань з великих даних запропоновано технологію, яка дозволяє отримувати із статистичних наборів даних логічні

залежності та формувати на їх основі логічні зв'язки та правила, за якими ці дані будуть відсортовувати та обробляти.

Обробка даних засобами запропонованої технології складається з таких етапів: очистки даних, нормалізації, кластеризації, побудови нечітких логічних правил та їх перевірки засобами математичного апарату метаграфів.

Під час очистки отримані статистичні дані аналізують на наявність групи помилок, таких як: відсутні значення, значення які є занадто великими або занадто маленькими (пікові значення) для певної галузі (наприклад, температура повітря 98 градусів за Цельсієм), значення, які є некоректними (наприклад, текстове значення серед цифрових). Більшість невалідних значень замінюють середніми значеннями, взятими з інших елементів таблиці, яка представляє зібрані статистичні дані.

Для процесу кластеризації застосовано математичний апарат нечіткої логіки, суть алгоритмів якого полягає у використанні інтелектуального способу міркувань, який спирається на природну особливість міркування людини та не може бути представлений традиційним математичним апаратом.

Нечітка логіка призначена для формалізації людських здібностей до неточних або наближених міркувань, які дозволяють більш адекватно описувати ситуації з невизначеністю. Це положення дозволяє побудувати логічну систему, в якій можна робити судження з невизначеністю і оцінювати ступінь істинності висловлювань. Одним з понять нечіткої логіки є поняття елементарного нечіткого висловлювання. Елементарне нечітке висловлювання представляє собою розповідне речення, яке є або істинним або хибним відносно певного факту із визначенням різного ступеня впевненості.

Відштовхуючись від теорії множин, елемент є здатним або належати до певної множини, або не належати до неї. Саме поняття множин є важливим у багатьох математичних теоріях, проте не розглядає такі випадки, коли ситуація проста і легко зрозуміла. Теорія нечітких множин базується на частковій приналежності до множини.

Бази правил нечіткої логіки є найчастіше використовуваними інструментами у прикладних програмних середовищах. Як правило, застосовують правила типу: IF «умова» THEN «результат». Бази, в яких зберігають правила нечіткої логіки ґрунтуються на базі знань. Разом з тим, має місце вплив людини під час формування таких правил, наприклад, один експерт або ціла група експертів задають для кожного аргументу ступінь приналежності (від 0 до 1), при цьому для них буде доцільніше обмежити розгляд параметрів тільки тими значеннями величин, які мають саме найбільшу важливість в рамках розв’язувальної задачі, а вже пізніше функція приналежності (шляхом багаторазових досліджень та перевірок) може бути уточнена. Головною проблемою під час формування нечіткого логічного виводу є те, що база правил має бути задана експертом заздалегідь і виключення експерта із цього процесу наразі не є можливим.

Кластерний аналіз - це пошуково-розпізнавальний процес, який використовується у великій кількості сфер людського життя. Кластеризація використовується у будь яких випадках, коли необхідно розмежувати купу інформації на певні групи або розділи, які в подальшому будуть цілком придатні для обробки. В останні роки процес кластерного аналізу широко використовують у Data Mining (інтелектуальний аналіз даних) як один з основних методів.

Мета кластерного аналізу полягає у призначені та віднесені об'єктів в однорідні набори даних, тобто так звані кластери. Призначення об’єктів проходить так, щоб об’єкти були схожими один на одного в одному кластері, а в інших відрізнялися. Спосіб узагальнення

спостережуваних даних у кластерах визначається на основі статистичної інформації, яка може зберігатися в таблицях баз даних або в файлах, оскільки на початку дослідження немає попередніх знань. Процеси різних алгоритмів кластеризації можна широко класифікувати так:

- Роздільні алгоритми, в яких кластери визначаються швидко. Об’єкти даних діляться на декілька розділів, кожен розділ представляє з себе кластер, який відповідає таким вимогам: кожна група має містити принаймні один об'єкт, кожен об'єкт має належати точно до однієї групи.
- Ієрархічні алгоритми організовують дані ієрархічним способом, залежно від їх близькості розташування. Початковий кластер поступово ділиться на декілька кластерів в залежності від наявної ієрархії. Цей процес буде продовжуватися до досягнення необхідного критерія зупинки (наприклад, певна кількість кластерів). Ієрархічні алгоритми мають і недолік, який полягає у тому, що після розділення даних на кластери їх неможливо повернути у попередній стан.
- Алгоритми на основі щільності поділяють дані, взявши за основу їх області щільності, зв’язності та кордону. Кластер, який було визначено як пов’язаний щільний компонент, здатний рости в будь-якому напрямку, до якого веде щільність.
- Алгоритми на основі сітки ділять простір об’єктів даних на сітки. Основна перевага такого процесу полягає у швидкості процесу обробки, оскільки він проходить через набір даних тільки один раз для обчислення статистичних значень для сіток. Зібрані дані сітки роблять методи кластеризації на основі сітки незалежними від кількості об’єктів даних, які використовують єдину сітку для збору статистичних даних, а потім виконують кластеризацію в сітці, а не безпосередньо в базі даних. Продуктивність методу на основі сітки залежить від розміру сітки, який зазвичай набагато менше розміру бази даних.

Для побудови нечітких логічних правил із очищених наборів даних в даному досліджені застосовано такі методи кластеризації: K-Means та FCM.

Побудова нечітких логічних правил має бути як можна точнішою, через складну природу великих даних. Тому важливим елементом є правильність кластеризації. Якщо на фінальному етапі кластеризації центри кластерів буде знайдено не правильно, то дані не правильно розділяться, через це буде похибка в побудові функцій приналежності. Для кластеризації методом K-Means та FCM потрібно скористатися алгоритмами первинної ініціалізації центрів кластерів та знаходження правильної кількості центрів кластерів. В модифікованому методі побудови нечітких логічних правил пропонується використати алгоритми kmeans++ для первинної ініціалізації та алгоритм «Ліктя» для знаходження правильної кількості кластерів в блоці кластеризації.

Запропонований метод працює тільки з числовими даними, які представлено в табличній формі, тому шо в нечіткій логіці на етапі фазифікації краще оперувати з числовими даними, для визначення нечіткого значення терму лінгвістичної змінної. В такій структурі даних, кожен стовпчик це властивість об’єкта, а кожен рядок і є об’єктом.

Якщо переходити до нечіткої моделі, то стовпці в таблиці розміщені таким чином, щоб перші *N*-стовпців були умовою, тобто антицедентом, що складає ліву частину нечіткого логічного правила. А останній стовпець є консиквентом, або нечітким логічним висновком, що знаходиться в правій частині нечіткого логічного виводу.

Правило має вид ЯКЩО…ТА….ТО. Результат нечіткого правила є комбінацією пропозицій об'єднаних операторами ТА. Інструкція «АБО» не використовується для

формування пропозицій результату, тому що вона вносить неоднозначність у правило (при цьому для виявлення коректної обробки необхідна додаткова експертиза).

Для того, щоб перейти до нечіткості потрібно для кожної лінгвістичної змінної знайти функції приналежності. Функція приналежності – це функція, що має значення від 0 до 1 по осі ординат, а по осі абсцис числові значення даного терма. Функції приналежності бувають різного виду, трикутні, трапецеоподібні або Гаусса [3, 1].

Після етапу побудови функцій приналежності кожен об'єкт із вхідної вибірки, створить окреме нове правило вигляду Якщо .. та … то [2-4]. Перетворення числових значень в терми лінгвістичних змінних відбувається на етапі фаззифікації. Спочатку визначається значення функції Гаусса в точці, для кожної функції приналежності, потім визначається найбільше значення із всіх значень, яке свідчить, що дана точка відноситься до нечіткої множини (до відповідного терма). Таким чином, для кожного числового значення рядка (характеристики об'єкта) знаходиться відповідна терм-множина, яка і буде нечітким логічним правилом.

Нижче сформовано набір правил вигляду Якщо … та .. то:

Якщо FR –> low та TR –> low то EN –> low

Якщо FR –> high та TR –> low то EN –> high

Якщо FR –> high та TR –> low то EN –> low

Якщо FR –> middle та TR –> low то EN –> high

На даному кроці перетворення даних в терми лінгвістичних змінних формуються конфліктуючі і дублюючі правила через різну природу даних або через те, що дані при кластеризації не правильно розподілилися між кластерами на границях розподілу кластерів. [2, 3,4]..

Модифікований метод побудови нечітких логічних правил для великих даних. Так як в даний метод добавлено вище зазначені алгоритми, то алгоритмічна та часова складність методу зростає. Проте більш важливим є правильність і точність побудови нечітких логічних правил. В модифікованому методі використано алгоритми kmeans++ для первинної ініціалізації та алгоритм «Ліктя» для знаходження правильної кількості кластерів в блоці кластеризації (рис.1).

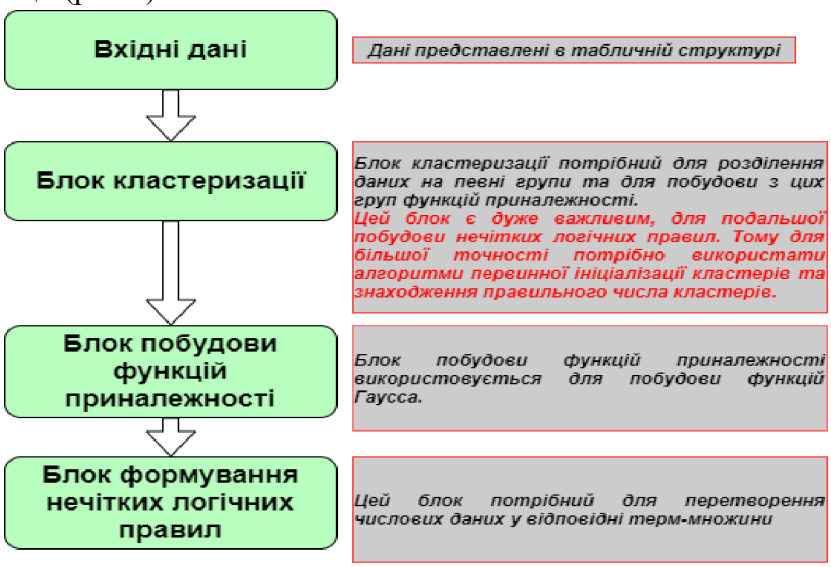

Рис. 1 Модифікований метод побудови нечітких логічних правил на базі статистичних даних

На рис.1 показані блоки з яких складається метод, а також показано червоним кольором, методи які було добавлено в блок кластеризації.

Таким чином, в даній статті проаналізовано математичний апарат нечіткої логіки та методи кластеризації, які використовується для побудови нечітких логічних правил.

Показано, як із числових статистичних даних будується функція приналежності, методами кластеризації.

В досліджені представлено механізм побудови нечітких логічних правил із статистичних числових даних. Даний механізм переводить кожне числове значення у терм-множину відповідної нечіткої змінної.

Запропонований підхід щодо отримання логічних залежностей із наборів статистичних даних реалізовано у проекті “Fuzzy logic”, який дозволяє мінімізувати вплив людини на обробку статистичних даних. В результаті обробки завантаженої інформації користувач отримує відсортовані та відкориговані дані, які можна використовувати у необхідних напрямках аналізу.

Технологія, на якій засновано проект “Fuzzy logic”, дозволяє обробляти статистичні дані та отримувати логічні залежності у вигляді нечітких логічних правил у реальному часі.

Запропоновані модифікації алгоритмів кластеризації збільшують точність побудови нечітких логічних правил за рахунок алгоритмів первинної ініціалізації та знаходження правильного числа кластерів.

Список літератури

1. Самсонов М. Ю. Интернет вещей в умном городе / М. Ю. Самсонов, А. Ю., Гребешков, А.В. Росляков, С.В. Ваняшин // ИнформКурьер-Связь, 2013. – № 10. – с. 58-61.
2. Globa L.S. INTELLIGENT SUPPORT SYSTEM FOR E-HEALTH / L.S. Globa, I.O. Ishchenko, A.G. Zakharchuk // 20th International Multi-Conference on Advanced Computer Systems Międzyzdroje, Poland – 2016. – p. 8.
3. Глоба Л.С., Бугаєнко Ю.М., Ляшенко А.В., Гребініченко М.В. Analysis of clustering algorithms for use in the universal data processing system // Міжнародна науково-технічна конференція OSTIS 2020 – с. 101–104.
4. Пегат А. Самоорганизующиеся и самонастраивающиеся нечеткие модели/ Анджей Пегат // Нечеткое моделирование и управление / Анджей Пегат., 2015. – (3-е издание). – с. 506–520.