УДК 621.39
© В. Періг[1], 2011

ПІДХОДИ ДО КЛАСТЕРНОГО АНАЛІЗУ ІНФОРМАЦІЙНО-ТЕЛЕКОМУНІКАЦІЙНИХ СИСТЕМ

Розглянуто сучасний стан кластерного підходу як інструменту для аналізу інформаційно-телекомунікаційних систем при розв'язуванні слабоформалізованих задач. Запропоновано методику застосування нейромереж для вирішення завдань кластеризації інформації. Висвітлено сучасний стан кластерного аналізу на основі порівняльного обговорення численних алгоритмів щодо можливості їх застосування для аналізу ІТС.

The modern state of cluster approach is in-process considered as to the instrument for the analysis of the informative-telecommunication systems at uniting of poorly formalized tasks. Methodology of application of neuron networks is offered for the decision of tasks to the clusterization of information. The modern state of cluster analysis is reflected on the basis of comparative discussion of numerous algorithms in relation to possibility of their application for the analysis of ITC.

1. ВСТУП

Останнім часом у різних галузях людської діяльності для автоматизації вирішення різноманітних інформаційних завдань використовують інформаційно-телекомунікаційні системи (ІТС). Від їхньої працездатності у багатьох випадках залежить можливість вирішення задач. Тому при виході ІТС з ладу необхідно оперативно діагностувати причини непрацездатності. Для вирішення цієї проблеми потрцібно провести аналіз (класифікацію стану) ІТС по сукупності параметрів, які описують її стан.

Одним із підходів до вирішення завдання класифікації об'єктів по сукупності ознак, які їх описують, є кластерний аналіз.

Головною метою кластерного аналізу є визначення множин схожих об'єктів у вибірці. Відсутність жорсткого визначення вимог до об'єктів класифікації у кластерному аналізі робить його універсальним інструментом. Саме тому спектр його прикладних застосувань є надзвичайно широким. Однак, спроби розробити універсальні підходи до застосування кластерного аналізу призвели до появи значної кількості методів [1, 2], застосування яких, в окремих випадках, ускладнює однозначне використання і коректну інтерпретацію результатів. Особливо це стосується

[1] Тернопільський національний економічний університет

тих завданнях, коли неможливо жорстко класифікувати об'єкти. Більшість методів кластерного аналізу потребує чіткого визначення однорідності об'єктів [1]. Однак, деякі задачі слабоформалізовані. Однією з них є аналіз інформаційно-телекомунікаційних систем.

Для опису ІТС використовують сукупність показників, окремі з яких не чітко визначені. При цьому, ґрунтуючись на сукупності показників загалом, неможливо однозначно функціонально встановити належність окремих об'єктів (ІТС) до кластерів. Достатньо складно також коректно визначити метрику для чіткого розмежування кластерів ІТС.

Тому актуальним на сьогоднішній день є дослідження методів кластеризації щодо можливості їх застосування для аналізу ІТС при розв'язуванні слабоформалізованих завдань. Метою статті є розробка нових підходів до аналізу інформаційно-телекомунікаційних систем.

2. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Незалежно від предмета, в межах якого проводиться класифікація, кластерний аналіз передбачає вирішення послідовності таких типових задач [1]: відбір вибірки для кластеризації; визначення множини ознак (параметрів), за якими проводиться кластеризація; вибір міри подібності об'єктів; застосування одного з методів кластерного аналізу для створення груп подібних об'єктів (кластерів); перевірка достовірності результатів.

Нехай X – множина об'єктів, які необхідно кластеризувати, Y – множина кластерів, X^n – скінченна множина (вибірка) об'єктів, на яких проводиться навчання ($X^n \subset X$).

На множині X задана метрика – функція визначення відстані між об'єктами. Кластеризація полягає в розбитті множини X на підмножини (кластери), які не перетинаються, таким чином, щоб кожен кластер містив об'єкти, близькі за метрикою. Об'єкти з різних кластерів повинні суттєво відрізнятись (знаходитись на значній відстані один від одного).

При цьому кожному об'єкту $x_i \in X^n$ ставиться у відповідність кластер $y_i \in Y$. Головне завдання кластеризації полягає у визначенні відображення

$$\alpha: X \to Y, \qquad (1)$$

за яким кожному об'єкту $x_i \in X$ ставиться у відповідність кластер $y_j \in Y$. В окремих випадках множина Y може бути визначеною завчасно. Однак, зазвичай, її визначення проводять при кластеризації. У цьому випадку необхідно знайти потужність множини кластерів і об'єкти (кластери), які до неї входять.

Проведемо аналіз існуючих підходів до вирішення задачі кластеризації інформації.

Перший підхід ґрунтується на використанні метода, який отримав назву K-means (к-середніх). Цей метод є найбільш популярними при вирішенні широкого кола прикладних задач кластеризації інформації. Особливість методу K-means полягає в тому, що кількість кластерів повинна бути відомою апріорі. При проведенні кластеризації мінімізується дисперсія на точках кожного кластера.

Основним положенням методу K-means є нове обчислення на кожній ітерації центра мас для кожного кластера, отриманого на попередньому кроці. Далі проводиться повторне розбиття на кластери відповідно до того, який з нових центрів є ближчим за обраною метрикою. Коли дві наступні ітерації не призводять до зміни кластерів, робота алгоритму кластеризації припиняється. Початкові центри кластерів і відповідно початкова кластеризація вибираються випадковим чином.

Основними недоліками методу K-means щодо застосування при вирішенні задачі аналізу ІТС є необхідність визначення метрики; випадковий вибір початкових кластерів; кількість кластерів потрібно задавати завчасно.

Можна модифікувати метод, визначивши в конкретному його прикладному застосуванні раціональне початкове розбиття, що дасть можливість суттєво підвищити його ефективність.

Оскільки при вирішенні задачі аналізу інформаційно-телекомунікаційних систем не завжди є можливим визначення метрики, використання одного з основних методів кластерного аналізу – K-means не є доцільним.

Інша група методів – графові методи кластеризації. В основу цих методів покладено використання біграфів для побудови кластерів. Біграф – це граф, який ділить множину на дві підмножини, що не перетинаються. Однак, при використанні графових методів для застосування біграфів необхідно визначати відношення між об’єктами, які підлягають кластеризації. Проте, їх визначення у випадку класифікації інформаційно-телекомунікаційних систем складно реалізувати.

Ще один відомий метод кластеризації – FOREL (FOREL2). В основу алгоритму, який реалізує цей метод, покладено використання принципу компактності. Близьким за змістом об’єктам у просторі ознак відповідають окремі множини (кластери). Однак, для визначення близькості необхідно на просторі ознак визначити метрику, що у випадку кластеризації інформаційно-телекомунікаційних систем не завжди є можливим.

Провівши аналіз ряду методів, які використовують для кластеризації інформації, можна зробити висновок, що вони, зазвичай, ґрунтуються на визначенні метрики в тому чи іншому вигляді. Відстань між об’єктами використовують для визначення областей близьких об’єктів

(кластерів). Однак, в окремих завданнях, зокрема в задачі кластеризації інформаційно-телекомунікаційних систем, визначення метрики є складним і не завжди однозначним. Тому в таких випадках доцільним є дослідження підходів, що не потребують явного визначення метрики.

Для проведення кластеризації без визначення метрики на просторі ознак об'єктів можна застосовувати методи штучного інтелекту, зокрема нейронні мережі. Для вирішення цього завдання традиційним є застосування нейронних мереж Кохонена [3].

Нейромережа Кохонена – нейронна мережа, розроблена Тойво Кохоненом на початку 1980-х років. Вона принципово відрізняється від багатьох інших нейромереж. Основна її особливість полягає у використанні неконтрольованого навчання. Вектор вхідних сигналів x містить параметри, які описують об'єкт („інформаційно-телекомунікаційну систему"). Кількість нейронів у шарі Кохонена відповідає кількості кластерів. Після обробки нейромережею вхідних даних лише один із нейронів має активний сигнал на виході (1). Саме до кластера, який відповідає цьому нейрону, і належить ІТС, характеристики якого подавались у вхідному сигналі.

Однак, при використанні нейромережі Кохонена виникає проблема лінійної подільності простору ознак. Тому при вирішенні задачі кластеризації інформаційно-телекомунікаційних систем можливим є також застосування більш складних за структурою нейромереж.

Проте, окремі характеристики ІТС можуть бути визначені нечітко, тому виникає необхідність адаптувати такі нейромережі до обробки вхідної нечітко заданої інформації.

3. ВИСНОВОК

На основі проведеного аналізу існуючих підходів до кластеризації інформації було встановлено, що більшість класичних методів потребує функціонального визначення відображення (1) або, принаймні, метрики в просторі ознак, що при вирішенні задачі кластеризації ІТС не завжди є можливим. Тому для аналізу ІТС доцільно за основу вибрати метод кластеризації із застосуванням нейромережі. Однак, для врахування нечіткості окремих параметрів ІТС необхідно адаптувати існуючу нейромережу.

1. Олдендерфер М. С. Кластерный анализ. Факторный, дискриминантный и кластерный анализ / Олдендерфер М. С., Блэшфилд Р. К. : пер. с англ.; Под. ред. И. С. Енюкова. — М. : Финансы и статистика, 1989. — 215 с. 2. The EM algorithm // The Elements of Statistical Learning. — New York: Springer, 2001. — P. 236–243. 3. Kohonen T. (1988), Learning Vector Quantization, Neural Networks, 1 (suppl. 1). – P. 303.