УДК 681.518:004.93.1

Н.В. Лисак[1]
С.О. Петров[2]

ГІБРИДНИЙ АЛГОРИТМ КЛАСТЕР-АНАЛІЗУ ДЛЯ ФОРМУВАННЯ АПРІОРНОГО РОЗБИТТЯ ПРОСТОРУ ОЗНАК НА КЛАСИ ЗНАНЬ В СИСТЕМАХ ДИСТАНЦІЙНОГО НАВЧАННЯ

[1] Вінницький національний технічний університет
[2] Сумський державний університет

Анотація

В роботі запропоновано модифікація алгоритму k-means ідея вдосконалення якого полягає у комбінованому використанні критерію оцінки помилки кластеризації та інформаційного критерію функціональної ефективності, що визначає рівень достовірності побудованих вирішальних правил визначення належності реалізацій до певного класу знань. При цьому використання комбінованого статистичного та інформаційного підходів дозволило включити такий параметр кластеризації як кількість кластерів в інтеграційну оптимізаційну процедуру та базуючись на природній структурі розподілення векторів реалізацій результатів тестування слухачів в N-вимірному просторі ознак розпізнавання дозволило знайти оптимальні геометричні параметрі контейнерів класів що характеризують рівні знань студенів в системах дистанційного навчання.

Ключові слова: *кластеризація, k-means, критерій функціональної ефективності, критерій оцінки помилки кластеризації, системи дистанційного навчання.*

Abstract

In the work we introduce a modification of the k-means algorithm, the idea of improvement consider in combined use of evaluation clustering errors criteria and test the functional efficiency, which determines the construction of decision rules accuracy for determining appurtenance to some implementations of the class of knowledge. Simultaneous use of statistical and informational approach allowed to include an important parameter for the clustering algorithms - the number of clusters in an iterative optimization process and on the other hand having a priori information about the distribution of N-dimensional vectors realizable test results of students' knowledge to determine the optimal geometric parameters of students knowlages container in scope of distance learning systems.

При інформаційному аналізі і синтезі адаптивної системи керування дистанційним навчанням (СКДН) на етапі навчання системи виникає необхідність формування апріорного розбиття простору ознак розпізнавання на класи, яке у процесі побудови вирішальних правил корегується деяким оптимальним способом [1]. Фактично, таку задачу розв'язує безпосередньо викладач, відносячи результати тестування до відповідної суб'єктивної та інтуїтивної оціночної шкали. Але оскільки при швидкому зростанні кількості слухачів, набуття популярності заочно-дистанційної форми навчання, спеціалізованих навчальних сертифікованих курсів, збереження традиційної системи оцінювання знань вимагає постійного збільшення матеріальних витрат, підвищення тиску на професорсько-викладацький склад, то задача розробки алгоритмів машинного оцінювання знань студентів за методами сучасних прогресивних інтелектуальних технологій є актуальною [2].

Окрім цього, у задачах з невеликою кількістю об'єктів, де більш важливим є аналіз структури даних, а також існує окрема проблема визначення кількості кластерів, використовують ієрархічні методи такі як: метод ближнього сусіда (single linkage), метод дальнього сусіда (complete linkage), метод середнього зв'язку (pair group average), центроїдний метод (метод медіан зв'язку) [3,4].

Всі відомі методи кластер-аналізу поєднуються такими властивостями:

застосування дистанційних критеріїв схожості (евклідова відстань, зважена евклідова відстань та відстань Хемінга.), які є частинними випадками узагальненої махаланобісової метрики [5]

Але до теперішнього часу теорія автоматичної класифікації не дає відповіді на такі запитання:

На скільки вдалим є побудоване на етапі навчання СКДН розбиття простору ознак на класи, що визначає вирішальні правила.

Чи є оптимальним (тут і далі в інформаційному розумінні) число класів розпізнавання?

При дослідженні цих важливих питань, як правило, здійснюється розвідувальний аналіз результатів тестування шляхом оцінки їх статистичних характеристик і емпіричних закономірностей [6]. При цьому для оцінки апріорного розбиття часто використовують таку величину, як

внутрікласовий розкид [7].

Тому, наприклад, у працях [7,8] запропоновано підхід, що ґрунтується на виявленні у емпіричних даних об'єктивно існуючої функціональної закономірності. Такий підхід має ряд недоліків:

висока обчислювальна трудомісткість, пов'язана із необхідністю оброблення великих масивів емпіричних даних;

наявність апріорно чіткого розбиття, що не є характерним для більшості практичних задач контролю та керування;

лінійній вигляд функції регресії, яка не у всіх випадках адекватно описує зв'язок.

Таким чином, аналіз існуючих сучасних методів кластер-аналізу дозволяє зробити такі два основні висновки:

а) методи кластер-аналізу, що виключають процес машинного навчання, характеризуються низькою достовірністю прийняття рішень відносно методів із навчанням;

б) відомі методи кластер-аналізу, що навчаються, не дозволяють побудувати безпомилковий за навчальною матрицею класифікатор, оскільки ігнорують перетин класів розпізнавання, тобто носять модельний характер.

Однією із перспективних технологій аналізу та синтезу адаптивних систем керування слабо формалізованими процесами є інформаційно-екстремальна інтелектуальна технологія (ІЕІТ), що ґрунтується на реалізації принципу максимізації інформаційної спроможності системи шляхом введення в процесі оптимізації просторово-часових параметрів функціонування додаткових інформаційних обмежень з метою побудови в режимі навчання безпомилкового за навчальною матрицею класифікатора [9].

У роботі розглядається алгоритм побудови у рамках ІЕІТ апріорного нечіткого розбиття класів розпізнавання (рівнів знань) студентів заочної і дистанційної форм навчання, що дозволяє сформувати вхідну навчальну матрицю для адаптивної СКДН.

Таким чином, розроблено модифікацію алгоритму k-means алгоритму та виконано дослідження його параметрів щодо кластеризації вхідних даних за умов їх апріорного природного впорядкування. З метою оцінки якості кластеризації оцінювання рівня знань студентів в системі дистацнійного навчання разом з КФЕ було використано розроблений новий критерій, який дозволив оцінити результативність кластерного аналізу в залежності від вхідних параметрів алгоритму таких як: прогнозована апостеріорна похибка прийняття рішення, кількість контейнерів класів розпізнавання, розмірність простору ознак розпізнаванняю Окремо досліджено зв'язок параметрів алгоритму з КФЕ, який показав, що в залежності від розмірності простору ознак стає можливим суттєво зменшення похибки результатів кластеризації вхідних даних, та дозволяє досягти асимптотичного максимуму КФЕ системи вцілому. З використанням запропанового гібридного підходу кластер-аналізу стає можливим зменшувати апостеріорні помилки на етапі функціонування СКДН в режимі екзамену за рахунок автоматичного визначення параметрів алгоритмів кластеризації.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Петров С.А. Категориально-информационная модель адаптивной системы непрерывного обучения / С.А. Петров // Управляющие системы и машины.–2009.-№2.– С.48-51.

2. Довбиш А.С. Машинна оцінка знань студентів у системах керування дистанційним навчанням / А.С. Довбиш, В.О. Любчак, С.О. Петров // Вісник Сумського державного університету. Серія «Технічні науки».– 2007.– №1.– С. 167-178.

3. Ким, Дж.-О. Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким, Ч.У. Миллер, У.Р. Клекк и др. – М.: Финансы и статистика, 1989. – 215 с.

4. Jain A. K., Murty M. N., Flynn P. J. Dataclustering: a review / A. K.Jain, M. N. Murty, P. J. Flynn // ACM Computeing Surveys(CSUR).–1999 – Volume 31 Issue 3 –69 p.

5. Айвазян С.А., Бухштабер В.М. Прикладная статистика. Классификация и снижение размерности. М.–Финансы и статистика. – 1989 г 300-310с.

6. Браверман Э.М., Мучник И.Б. Структурные методы обработки эмпирических данных.–М.:Наука.Физматлит,–464 с.

7. Алехин Е. И. Многомерные статистические методы / Е. И. Алехин.– Орел: Издательский центр ГОУ ВПО ОГУ, 2007 – 37с.

8. Бабак О.В., Касанов А.С. Алгоритм решения некоторых задач кластерного анализа. УСИМ: Управляющие системы и машины, 2001, №6 25-30 с.

9. Петров С.О. Вплив структури простору ознак розпізнавання в системах підтримки прийняття рішень / С.О. Петров // «Інтернет-Освіта-Наука –2010»: Сьома міжнар. конф. ІОН-2010: 28 вер.-3 жов. 2010 р.: тези доп.– Вінниця: Вінницький Національний технічний університет, 2010.– С.71-72.

Лисак Наталія Володимирівна – к.т.н., доцент кафедри менеджменту та безпеки інформаційнихсистем, Вінницький національний технічний університет, м. Вінниця

Петров Сергій Олександрович – к.т.н., старший викладач комп'ютерних наук, Сумський державний університет, Суми