УДК 31.311.2
Ханін[1] О.Г., Лотиш[2] В.В., Гуменюк[2] П.О.
[1]Східноєвропейський національний університет імені Лесі Українки
[2]Луцький національний технічний університет,

ІДЕНТИФІКАЦІЯ СИСТЕМ МАСОВОГО ОБСЛУГОВУВАННЯ З КОНТРОЛЬОВАНОЮ ЙМОВІРНІСТЮ ПОХИБКИ ТА ЇЇ ПРОГРАМНА РЕАЛІЗАЦІЯ

Ханін О.Г., Лотиш В.В., Гуменюк П.О. Ідентифікація систем масового обслуговування з контрольованою ймовірністю похибки та її програмна реалізація. Запропоновано метод кластерного аналізу даних, що мають довільний дискретний розподіл, який дозволяє на певному рівні значущості приймати статистично обґрунтовані рішення про належність об'єкта до певного кластера. Метод дозволяє у багатьох практичних задачах, зокрема задачах ідентифікації систем масового обслуговування швидко проводити статистично обґрунтований кластерний аналіз довільно розподілених даних із заздалегідь визначеною ймовірністю похибки кластеризації
Ключові слова: кластерний аналіз, системи масового обслуговування, χ2-розподіл, довірчий інтервал надійності, програмна реалізація, програмний додаток.

Khanin O.G., Lotysh V.V., Gumeniuk P.O. Identification of mass-service systems with controlled capacity and its program realization. The method of cluster analysis of data with arbitrary discrete distribution, which allows at a certain level of significance to make statistically substantiated decisions about the belonging of an object to a particular cluster, is proposed. The method allows many practical tasks, in particular, the problems of mass service identification, to quickly carry out a statistically valid cluster analysis of randomly distributed data with a predetermined probability of clustering error
Key words: cluster analysis, mass maintenance systems, χ2-distribution, confidence interval of reliability, software implementation, software application.

Ханин А.Г., Лотыш В.В., Гуменюк П.А. Идентификация систем массового обслуживания с контролируемой вероятность погрешности и ее программная реализация. Предложен метод кластерного анализа данных, имеющих произвольное дискретное распределение, который позволяет на определенном уровне значимости принимать статистически обоснованные решения о принадлежности объекта к определенному кластеру. Метод позволяет во многих практических задачах, в частности задачах идентификации систем массового обслуживания, быстро проводить статистически обоснованный кластерный анализ произвольно распределенных данных с заранее определенной вероятностью ошибки кластеризации.
Ключевые слова: кластерный анализ, системы массового обслуживания, χ2-распределение, доверительный интервал надежности, программная реализация, программное приложение.

У задачах автоматизованого управління часто виникає необхідність кластеризації даних для напрацювання однотипних алгоритмів управління, наприклад, задача ідентифікації кількох багатоканальних систем масового обслуговування. Існує чимало методів кластеризації [1, 2], деякі з них реалізовані в програмах обробки статистичних даних, таких, як SPSS Statistics або STATISTICA. Ці методи дають можливість вивчити структуру даних, але не роблять ніяких статистичних висновків.

Нами пропонується метод кластерного аналізу даних, що мають довільний дискретний розподіл, який дозволяє на певному рівні значущості приймати статистично обґрунтовані рішення про належність об'єкта до певного кластера.

В ході реалізації будь-якого методу багатовимірної кластеризації виникає необхідність деякого нормування даних. Запропонований метод передбачає попереднє нормування даних так, щоб кожен замір знаходився в межах від 0 до 1, а їх сума дорівнювала 1. Таким чином, буде відбуватися кластеризація об'єктів за статистичним розподілом певних ознак.

Наприклад, розглянемо результати дослідження завантаженості (за кількістю заявок на обслуговування за одиницю часу) сукупності m-канальних систем масового обслуговування. Нехай, для визначеності, *m = 4*.

Таблиця 1. **Розподіл завантаженості каналів систем масового обслуговування (СМО).**

СМО	Разом	Канал1	Канал2	Канал3	Канал4
СМО1	319	69	95	8	147
СМО2	**169**	**12**	**75**	**6**	**76**
……					
СМОk	284	33	51	17	183

Для кожної системи масового обслуговування отримаємо відносну частоту, поділивши кількість заявок на обслуговування, що надійшли в певний канал, на загальну кількість заявок, що надійшли в систему.

Таблиця 2. **Розподіл відносних частот завантаженості каналів СМО**

СМО	**Канал1**	**Канал2**	**Канал3**	**Канал4**
СМО1	0,22	0,30	0,02	0,46
СМО2	**0,07**	**0,44**	**0,04**	**0,45**
……				
СМО k	0,11	0,42	0,06	0,41

В основу методу кластеризації покладемо ідею порівняння емпіричних $\chi 2$-розподілів [3]. На першому етапі виберемо будь-яку СМО в якості еталонної, наприклад, СМО2 (в таблицях виділена жирним шрифтом). Побудуємо для еталонної системи по кожному каналу окремо довірчий інтервал надійності 95% для теоретичної ймовірності, що заявка, яка випадково надійшла в систему, надійде на обслуговування саме в даний її канал.

Обсяг вибірки по еталонній СМО (загальна кількість заявок, що надійшли) по таблиці 1 становить $n=169$. Розглянемо надходження випадкової заявки в 1-й канал СМО як «успіх», а інші - як «невдачу». Повторимо по-черзі цей процес для інших каналів. Тоді ми будемо мати справу з біноміальним розподілом для теоретичних ймовірностей, для яких легко побудувати двосторонній асимптотичний довірчий інтервал будь-якої надійності [4].

Так, вибіркова оцінка невідомого стандартного відхилення теоретичної ймовірності «успіху»

$$s_n = \sqrt{\frac{w_{усп}(1-w_{усп})}{n}} \approx 0{,}002,$$

де $w_{усп} \approx 0{,}07$ - відносна частота «успіху» (див. таблицю 2).

Тоді права границя довірчого інтервалу становить

$$w_{усп} + t_{0{,}95} \cdot s_n \approx 0{,}108,$$

де $t_{0{,}95} \approx 1{,}96$ – відповідний квантиль двостороннього розподілу Стьюдента з $n-1=168$ ступенями свободи, а ліва –

$$w_{усп} - t_{0{,}95} \cdot s_n \approx 0{,}032.$$

Тобто з надійністю 95% ймовірність p того, що випадкова заявка, яка надійшла в СМО2 потрапить на обслуговування в «Канал1», знаходиться в інтервалі (0,032; 0,108).

Так само побудуємо довірчі інтервали надійності 95% для теоретичних ймовірностей, які відповідають іншим каналам.

$\chi 2$ - відстань між емпіричним і теоретичним розподілом, знаходиться за формулою [5]

$$\chi^2 = \frac{1}{n} \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{p_i}, \quad (1)$$

де r – кількість груп, на які розбиті дані (в нашому випадку - кількість каналів, тобто $r=4$);

p_i – теоретичні ймовірності. У нашому випадку це ймовірність того, що випадкова заявка, яка надійшла в еталонну СМО, надійде на обслуговування в *і-ий* канал ($i=1,2,3,4$). Ці ймовірності нам невідомі, але ми встановили довірчі інтервали, в яких вони знаходяться (див. таблицю 3);

ν_i – вибіркові частоти кожної групи (в нашому випадку - кількість заявок, що надійшли в нееталонну СМО, які надійшли в її *і-ий* канал - таблиця 1);

$\underline{n}$ – обсяг вибірки (в нашому випадку n дорівнює загальній кількості заявок, що надійшли в нееталонну СМО - таблиця 1).

Таблиця 3. **Границі довірчих інтервалів для теоретичних ймовірностей поступлення на обслуговування до відповідних каналів еталонної СМО.**

Канали еталонної СМО	Ліва границя довірчого інтервалу	Права границя довірчого інтервалу
Канал1	0,032	0,108
Канал2	0,365	0,515
Канал3	0,010	0,070
Канал4	0,375	0,525

Таким чином, ми будуємо довірчі границі заданої надійності для теоретичного (еталонного) розподілу, за допомогою яких на першому етапі будемо порівнювати за критерієм $\chi 2$ емпіричні розподіли по інших СМО з розподілом по еталонній.

В один кластер з еталонною об'єднаємо СМО, емпіричні розподілу поступлення заявок в канали яких значуще не відрізняються за критерієм $\chi 2$ від теоретичного розподілу для еталонної СМО. Оскільки нам невідомі точні значення теоретичних ймовірностей pi еталонної СМО, замінимо їх на такі значення з довірчих відрізків (тобто з довірчих інтервалів, включаючи їх кінці), які зроблять значення $\chi 2$ у виразі (1) найменшим з можливих, тим самим при перевірці гіпотези про згоду розподілів ми мінімізуємо помилку 1-го роду. Зауважимо, що довірчі відрізки не повинні містити нульові значення. Якщо це трапилось, тобто значення відносних частот для деяких категорій занадто малі, варто об'єднати ці категорії з іншими. Якщо отриманий мінімум є більшим, ніж критичне значення, то розглянута СМО не входить в один кластер з еталонною, в іншому випадку СМО об'єднуються в один кластер. Після порівняння кожної з інших СМО з еталонною, закінчується формування першого кластера.

На наступному етапі в якості еталонної береться СМО, яка не потрапила в перший кластер, і аналогічним чином формується другий кластер. Процес кластеризації повторюється, поки кожна СМО не потрапить в певний кластер.

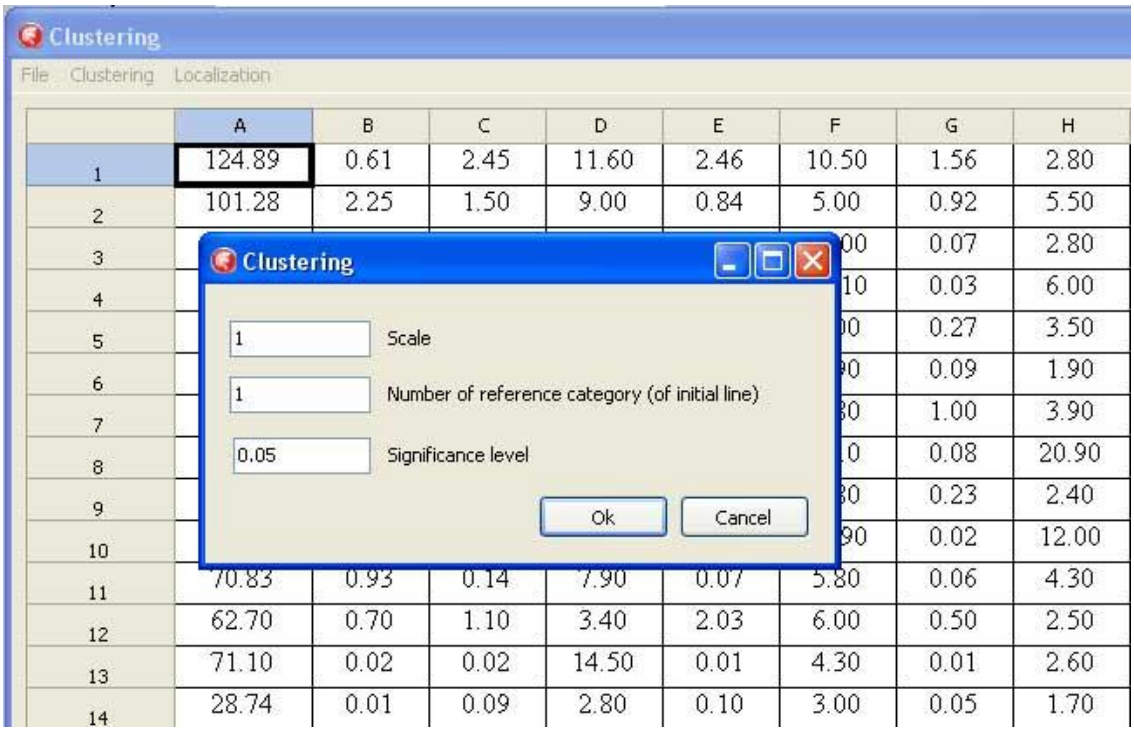
Програмна реалізація вищеописаного підходу виконана в середовищі програмування RAD Studio XE6. Отриманий програмний додаток має меню, що складається із зони роботи з файлами, зони кластеризації і зони налаштування мови інтерфейсу.

Дані для кластеризації можуть бути підготовлені як в Excel, так і введені безпосередньо в додатку.

Перед початком кластеризації необхідно в діалоговому режимі задати масштаб (коефіцієнт множення), номер еталонної категорії (в нашому випадку - номер СМО) і p-рівень значущості (Рис.1).

Результат розрахунку (кластеризація) виводиться на екран у вигляді ствпчикової діаграми, де по осі абсцис представлені номери кластерів, а по осі ординат - номери категорій (в нашому випадку - номери СМО), що входять в даний кластер (Рис.2).

Отримані результати можна експортувати в Excel і / або зберігати у вигляді графічного файлу.



	A	B	C	D	E	F	G	H
1	124.89	0.61	2.45	11.60	2.46	10.50	1.56	2.80
2	101.28	2.25	1.50	9.00	0.84	5.00	0.92	5.50
3						00	0.07	2.80
4						10	0.03	6.00
5						00	0.27	3.50
6						90	0.09	1.90
7						80	1.00	3.90
8						.0	0.08	20.90
9						80	0.23	2.40
10						90	0.02	12.00
11	70.83	0.93	0.14	7.90	0.07	5.80	0.06	4.30
12	62.70	0.70	1.10	3.40	2.03	6.00	0.50	2.50
13	71.10	0.02	0.02	14.50	0.01	4.30	0.01	2.60
14	28.74	0.01	0.09	2.80	0.10	3.00	0.05	1.70

Рис.1. **Приклад вводу параметрів кластеризації**

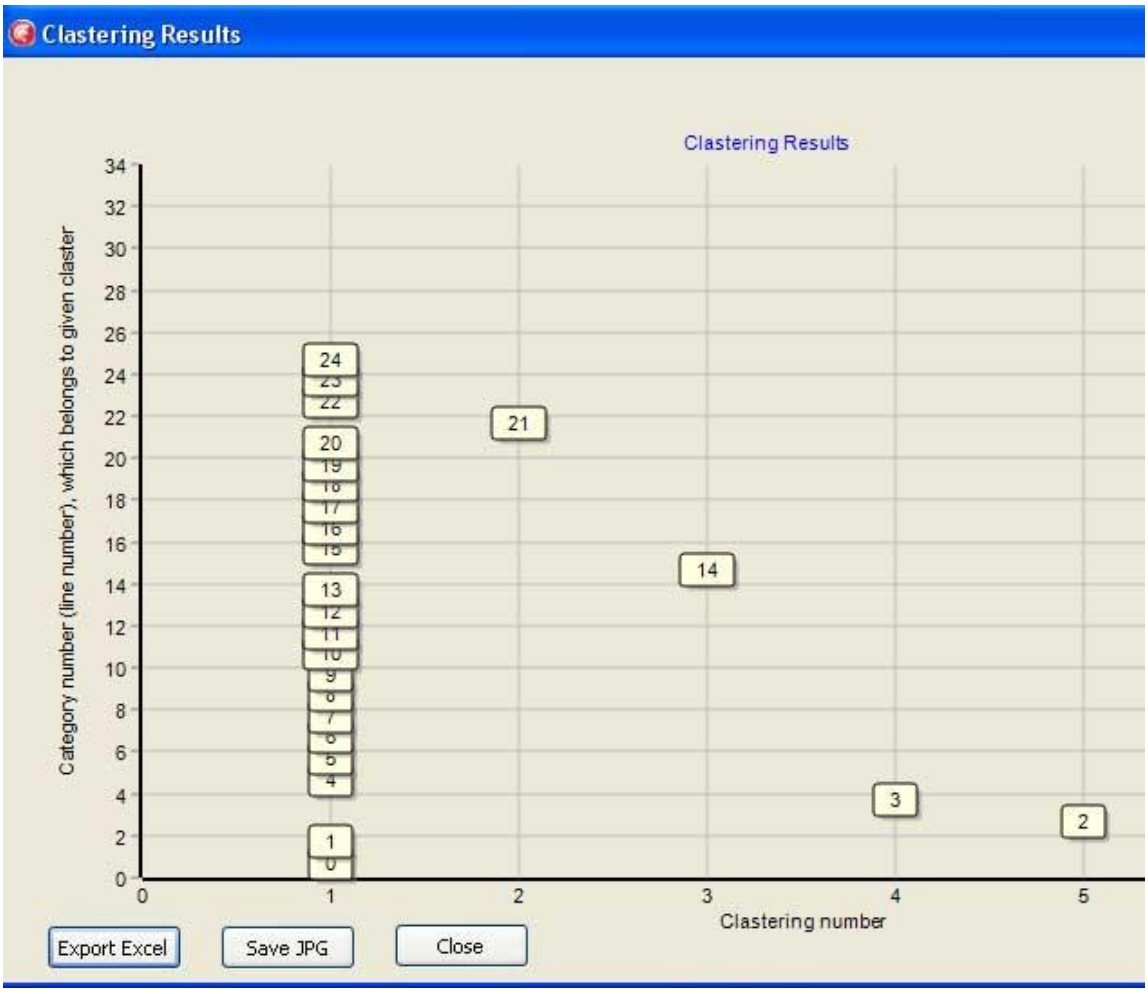


Рис 2. **Приклад виводу результатів кластеризації**

Таким чином, запропонований метод дозволяє у багатьох практичних задачах, зокрема задачах ідентифікації систем масового обслуговування, за допомогою критерію згоди *χ2* швидко проводити статистично обгрунтований кластерний аналіз довільно розподілених даних із заздалегідь визначеною ймовірністю похибки кластеризації.

1.Нейский, И. М. Классификация и сравнение методов кластеризации [Електронний ресурс] / И. М. Нейский. - Режим доступа: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf
2. Пилипчук, А. В. Организация фирменных торговосбытовых систем в агропромышленном комплексе Беларуси.- Минск: Ин-т системных исследований в АПК НАН Беларуси, 2011.- 178 с.
3.Ханін, О. Г. Методологічні особливості застосування критерію узгодженості χ^2 в практичних задачах економіки, соціології та маркетингу/О.Г.Ханін//Економічний аналіз.-2015.-Том 22,№ 1, с.67-70.
4. Сигел, Э. Практическая бизнес-статистика.- М.:Вильямс,2002.-1056 с.
5.Крамер, Г. Математические методы статистики.- М.:Мир,1976.-648 с.
6.Антонова И., Карих О. Оценка эффективности параллельных алгоритмов задачи сортировки данных. Промышленные АСУ и контроллеры., 2010., № 3., С. 23–25.
7.Коварцев А., Попова-Коварцева Д. Структурная оптимизация управляющего графа на основе алгоритма топологической сортировки. Программная инженерия., 2013., № 5., С. 31–36.
8.Кнут Д. Искусство программирования. Т.3. Сортировка и поиск /Д.Э. Кнут: пер. с англ.- 2-е изд.- М.: Издательский дом "Вильямс"., 2003.- 832с.
9.Мартынов В., Миронов В. Параллельные алгоритмы сортировки данных с использованием технологии MPI. Вестник Сыктывкарского университета. Серия 1: Математика. Механика. Информатика., 2012., № 16., С. 130–135.
10. Овчинникова И., Сахнова Т. Алгоритмы сортировки при решении задач по программированию. Информатика и образование., 2011, № 2., С. 53–56.
11. Самунь В. Сравнение работы алгоритмов сортировки, реализованных на языке Perl., 2007. стр. 21.