УДК 681.513.8

МЕТОД НЕЧІТКИХ K-СЕРЕДНІХ З ОБМЕЖЕНОЮ МАСОЮ РОБОЧОЇ ОБЛАСТІ ФОРМУВАННЯ КЛАСТЕРІВ ДОВІЛЬНОЇ ФОРМИ

Настенко Є. А., д.б.н., с.н.с.
e-mail : nastenko.e@gmail.com
Кафедра біомедичної кібернетики
Київський політехнічний інститут імені Ігоря Сікорського»
Київ, Україна
Уманець В. С., бакалавр
e-mail: 2_bytes@ukr.net
Факультет біомедичної інженерії
Київський політехнічний інститут імені Ігоря Сікорського»
Київ, Україна

Реферат. *Завдання визначення функціонального зв'язку між біофізичними показниками є складовою частиною вирішення актуальної проблеми пошуку оптимальних впливів на біологічний об'єкт і не вирішено на даний час в повній мірі. Однією з важливих задач в цій області є розбиття простору ознак на області (кластери), які відносяться до різних функціональних співвідношень, що зв'язують біофізичні показники, шукані кластери при цьому можуть мати довільну форму. Такі кластери назвемо функціональними, в роботі ставиться задача розробки методу виділення з вихідної вибірки даних кластерів довільної форми. Для вирішення поставленої задачі в роботі розглядається нечітка версія кластеризації для алгоритму k-середніх з обмеженою масою робочої області формування кластерів. Оцінка кількості кластерів проводиться за гістограмою частот, для визначення оптимальної кількості стовпців гістограми обґрунтовується застосування формули Скотта. Алгоритм дозволяє формувати кластери довільної конфігурації з отриманням значення міри належності об'єкта до кожного з кластерів. Ефективність алгоритму продемонстрована на прикладі кластеризації набору даних «Ірриси Фішера». Проведено порівняльне тестування: класичний алгоритм k-середніх, метод Варда та розроблений алгоритм. Результати, що одержано, дозволили віддати перевагу в задачі аналізу кластерів довільної форми розробленій в даній роботі версії нечіткого k-середніх з обмеженою масою робочої області формування кластерів. Розрахунок функції належності дозволяє отримати додаткову інформацію про структуру кластерних утворень, а також здійснити поправки результату кластеризації k-середніх з обмеженою масою, що особливо важливо для алгоритмів, що отримують результат кластеризації за один прохід. Відносно порівняння якісних результатів розробленого алгоритму та алгоритму Варда слід відмітити, що розроблений алгоритм має нижчу обчислювальну вартість так як не вимагає додаткової пам'яті для зберігання матриці відстаней та часу на її перерахунок. Крім того, розроблений алгоритм не має проблем, пов'язаних з розрізом дендрограми для отримання кластерів.*

Ключові слова: *кластеризація, k-середніх, міра належности, оцінка кількості кластерів, нечітка кластеризація.*

I. Вступ

Завдання визначення функціонального зв'язку між біофізичними показниками є складовою частиною актуального завдання пошуку оптимальних впливів на біологічний об'єкт і не вирішені в повній мірі в даний час. При цьому найбільш цікавими є результати, що адекватно представляють розбиття простору на області (кластери) які відносяться до різних функціональних співвідношень, що зв'язують біофізичні показники, що розглядаються, в даній області. Такі кластери логічно називати функціональними, а їх форма в загальному випадку може бути довільною. Для адекватного поділу вихідної сукупності на такі однорідні групи необхідне застосування нових інформаційних технологій.

II. Аналіз літературних даних та постановка проблеми

Одним з найбільш поширених підходів до кластеризації багатовимірних даних прийнято вважати методи з сімейства k-середніх. Однак коректне застосування класичної версії підходу конструктивно призначене для формування виключно багатовимірної сферичної форми кластерів. Дана проблема долається введенням обмеження на сумарну масу робочої області,

за допомогою якої визначається поточне значення центроїда кластера. Даний підхід реалізується однією з актуальних версій алгоритму [1]. Однак і дана версія алгоритму має ряд недоліків: необхідність задавати кількість груп перед проведенням кластеризації і відсутність механізму розрахунку міри належності до кластера. Відзначимо що ряд результатів, заснованих на інформаційній ентропії [2, 3] та дивергенції [2, 4], вирішують проблему оцінки кількості кластерів, але вони мають досить високу обчислювальну складність, тому бажано мати більш простий механізм отримання даної оцінки. Нижче розглядається нечітка модифікація алгоритму k-середніх з обмеженою масою робочої області формування кластерів і перевіряється його ефективність для задачі із заданою довільною формою кластерів. У подальших роботах передбачається дослідити можливості використання даного алгоритму для вирішення завдання виділення функціональних кластерів.

III. Мета і задачі дослідження

Метою є розробка версії методу k-середніх, що вирішує задачу розбиття вихідної вибірки з формуванням кластерів довільної форми.

Задачами дослідження є розробка версії алгоритму нечіткої кластеризації для методу k-середніх з обмеженою масою робочої області формування кластерів, введення в алгоритм механізму оцінки кількості кластерів, а також дослідження ефективності отриманого алгоритму на контрольному прикладі розбиття вибірки даних з кластерами довільної форми.

IV. Основна частина

Стандартний механізм алгоритму k-середніх без обмеження маси (кількості) об'єктів робочої області формування кластерів призводить до отримання, в граничному положенні, кластерів сферичної форми, які ототожнюються з ідеальною формою груп об'єктів. При цьому шлях центроїдів до граничного стану не є ні предметом аналізу алгоритму, ні конструктивним елементом, що формує кластер. Має значення лише стійкість граничного стану центроїда, що і визначає кінцевий результат кластеризації.

В роботі [1] механізм алгоритму k-середніх вперше був застосований для отримання кластерів не сферичної форми, причому в основі визначення форми одержуваного кластера лежить вже не граничне положення центроїда, а шлях який центроїд проходить до свого граничного стану. Зсув центроїда робочої області визначає тренд робочого положення кластера і фактично дозволяє алгоритму здійснювати розпізнавання його фрагментів. Однак при реалізації стандартного механізму k-середніх по мірі приєднання нових об'єктів в робочу область швидкість руху центроїда неухильно знижується. Це пояснюється неухильним зниженням впливу одиночного об'єкта, що приєднується, по відношенню до накопиченої раніше маси робочої області тому і вплив на тренд нового об'єкта стає незначним. Введення в [1] обмеження на масу робочої області формування кластера дозволило запропонувати механізм формування кластерів довільної форми і поширити метод k-середніх на загальний випадок завдання кластерного аналізу. Однак, як зазначалося вище, дана версія алгоритму може бути доцільно доповнена оцінкою кількості кластерів в даній вибірці даних і розрахунком міри належності об'єктів до найближчих кластерів.

A. Опис алгоритму кластеризації

Нехай кожен об’єкт вихідного масиву N спостережень описується m-мірним вектором $\{X_1, X_2, \ldots, X_m\}$ та може бути представлений у вигляді точки в просторі ознак розмірності m.

Алгоритм включає в себе наступні кроки:

1) Нормування даних

2) Ініціалізація центрів робочої області формування кластерів, що відбувається одним із наступних способів:

a. Найближчі до початку координат;

b. Найближчі до центру мас множини точок в просторі ознак;

c. На периферії множини точок в просторі ознак;

d. Рівномірно віддалені від центру з заданим кроком;

e. Найбільш віддалені від початку координат;

f. Вибрані з окремих міркувань;

g. Вибрані випадковим чином.

3) Далі вибирається об'єкт l і розраховується відстань від l до кожного з k_t центроїдів;

4) Об'єкт приєднується до того кластеру, відстань до якого найменша;

5) Відбувається перерахунок положення центроїда за такими формулами:

якщо $n_t < I_{\max}$, то

$$x_{c_t} = \frac{\sum_{i=1}^{n_t} x_i + x_l}{n_t} \quad (1)$$

якщо $n_t = I_{\max}$, то

$$x_{c_t} = \frac{\sum_{i=1}^{n_t} x_i + x_l - p \cdot o}{I_{\max} - p}, n_t = n_t - p \quad (2)$$

6) Де n_t – кількість точок в кластері з індексом t які використовуються для розрахунку положення центроїда, I_{max} – максимальна кількість таких точок, C_t – кластер з індексом t, сума $\sum_{i=1}^{n_t} x_i + x_l$ – накопичена інформація, а p-кількість умовних об'єктів o координати яких дорівнюють поточним координатам центроїда, та які забуваються при розрахунку поточного положення центроїду.

7) Як видно з формули (2), при наявності граничної кількості точок в кластері відбувається "забування" частини попередньо накопиченої інформації, що дозволяє контролювати переміщення центроїда в процесі кластеризації. Адекватний вибір параметрів забезпечує більш впорядкований рух при відтворенні функціональної залежності.

8) Якщо не обумовлено окремо, процедура завершується після перебору всіх N об'єктів вибірки.

Кластеризація об'єктів здійснюється в однопрохідному варіанті і кластери, що одержуються в результаті мають несферичну форму.

Описаний вище алгоритм було пропоновано в [1] проте для одержання адекватного результату він потребує завдання кількості кластерів. Для вирішення цієї проблеми було запропоновано наступний підхід. Проводиться побудова гістограми щільності розподілу для кожної з m змінних простору кластеризації. В результаті підрахунку кількості згущень, в яких групуються значення об'єктів по даній змінній, можна отримати оцінку кількості кластерів у вибірці, як найбільшу кількість локальних максимумів на гістограмі. Для вирішення питання оптимальної кількості стовпців гістограми можна використовувати формулу Скотта, формулу Фрідмана-Діаконіса або аналогічні їм. При реалізації алгоритму була використана формула Скота [5], за рахунок більш низької обчислювальної вартості в порівнянні з формулою Фрідмана-Діаконіса.

Розглянемо варіант алгоритму, що забезпечує розрахунок міри приналежності об'єкту. Розрахунок значення функції приналежності в нашому випадку некоректно проводити тим же шляхом, що використовується в алгоритмі «С-середніх», оскільки в «С-середніх» обчислення нового положення центроїда відбувається після накопичення інформації, а не в процесі його руху. У випадку, коли положення центроїда змінюється в процесі приєднання точок, розраховане значення функції належності втратить свою актуальність у зв'язку зі зміною положення центроїда. В такому випадку значення функції належності слід розраховувати вже після формування кластерів. Крім того, зміні підлягає механізм розрахунку міри належності, так як на відміну від класичної версії С-середніх (формування кластерів гіперсферичної форми) кластери, що одержуються будуть мати стрічкоподібну форму.

Можливо пропонувати наступні варіанти:

1) Використання середньої відстані від точки, що досліджується, до всіх інших точок кластера;

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\frac{1}{N_j - 1} \sum_{l=1}^{N_j} \left\| x_i - x_l^{(j)} \right\|}{\frac{1}{N_k} \sum_{l=1}^{N_k} \left\| x_i - x_l^{(k)} \right\|} \right)^{\frac{1}{m-1}}}$$

1) Використання відстані від досліджуваної точки до «сліду», що залишає центроїд що переміщується.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i - t_{c_j} \right\|}{\left\| x_i - t_{c_k} \right\|} \right)^{\frac{1}{m-1}}}$$

В. Перевірка роботи алгоритму на тестовий вибірці

Перевірка роботи алгоритму проводилася на наборі даних «Іриси Фішера» [6].

Набір даних «Іриси Фішера» містить 150 ірисів трьох видів, по 50 кожного виду. В результаті проведення оцінки кількості кластерів за допомогою розробленої процедури було встановлено, що в наборі даних присутні 3 кластери. При кластеризації набору даних «Іриси Фішера» був отриманий наступний результат (табл. 1).

Таблиця 1. **Результат кластеризації**

		Реальні			***Сума***
		0	***1***	***2***	
Результат	***0***	36	0	0	36
	1	14	50	0	64
	2	0	0	50	50
Сума		50	50	50	150
Доля розпізн., %		72	100	100	

Як можна бачити, результат кластеризації виявився близький до реально існуючих груп. Значення F_1 міри було отримано з використанням macro-averaging [7] і склало 0,92. Нижче вказані результати, отримані за допомогою класичного методу k-середніх (табл. 2) і ієрархічної кластеризації методом Варда (табл. 3).

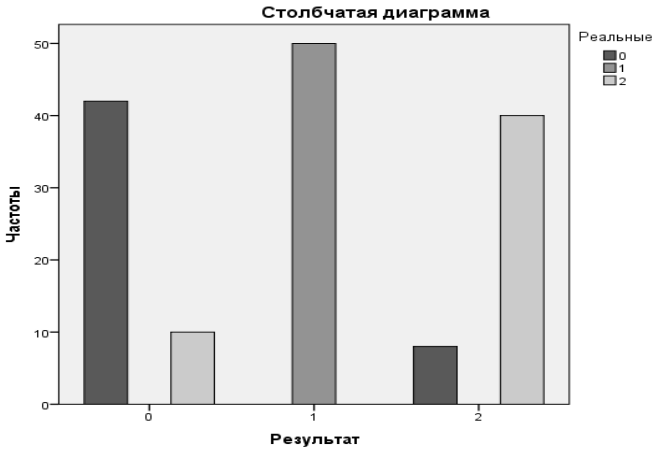

Рис. 1. Стовпчаста діаграма результату кластеризації методом k-середніх

Таблиця 2. **Результат кластеризації алгоритмом k-середніх**

		Реальні			***Сума***
		0	***1***	***2***	
Результат	***0***	42	0	10	52
	1	0	50	0	50
	2	8	0	40	48
Сума		50	50	50	150
Доля розпізн., %		84	100	80	

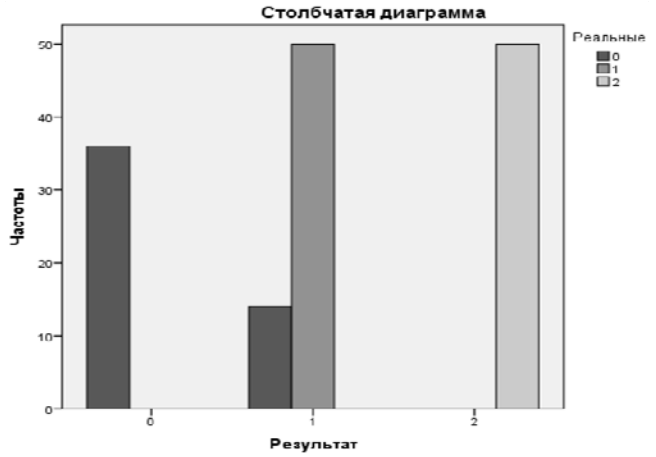

Рис. 2. Стовпчаста діаграма результату ієрархічної кластеризації методом Варда

Таблиця 3. **Результат ієрархічної кластеризації методом Варда**

		Реальні			***Сума***
		0	***1***	***2***	
Результат	***0***	33	0	0	33
	1	0	50	0	50
	2	17	0	50	67
Сума		50	50	50	150
Доля розпізн., %		66	100	100	

Як можна бачити з отриманого результату, ієрархічна кластеризація методом Варда породила кластери, схожі на ті, які були отримані за допомогою k-середніх з обмеженою масою робочої області. Деяка схожість результатів кластеризації є наслідком породження ієрархічними алгоритмами несферичних кластерів, в загальному випадку. Значення F1 міри для даного результату склало 0,90.

При тестуванні роботи алгоритму використовувався розрахунок функції належності з використанням мінімальної відстані до шляху центроїда робочої області алгоритму що забезпечило кращий результат кластеризації в порівнянні з використанням середньої відстані від досліджуваної точки до всіх інших точок кластера.

V. Висновки з дослідження і перспективи роботи

Тестування алгоритмів, що розглянуто в статті, дозволяє віддати перевагу в задачах аналізу кластерів довільної форми розробленій в даній роботі версії нечіткого k-середніх з обмеженою масою робочої області формування кластерів. Розрахунок функції належності дозволяє отримати додаткову інформацію про структуру кластерних утворень, а також здійснити поправки результату кластеризації k-середніх з обмеженою масою, що особливо важливо для алгоритмів, які отримують результат кластеризації за один прохід. Відносно близькості якісних результатів розробленого алгоритму і алгоритму Варда слід згадати, що розроблений алгоритм має нижчу обчислювальну вартість так як не вимагає додаткової пам'яті для зберігання матриці відстаней і часу на її перерахунок. Крім того, розроблений алгоритм не має проблем, пов'язаних з розрізом дендрограми для отримання кластерів.

СПИСОК ЛІТЕРАТУРИ

[1] E. Nastenko, «The use of Cluster Analysis for Partitioning,» *J. of Automation and Information Sciences,* pp. 77-83, 1996.

[2] Болдак, А.А.; Сухарев, Д.Л., «Определение количества кластеров в статистических данных,» *Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка,* p. 118–122, 2011.

[3] C. Shannon, «A Mathematical Theory of Communication», *Bell System Tech,* pp. 379- 423, 623-656, 1948.

[4] S. Kullback та R. Leibler, *The Annals of Mathematical Statistics,* p. 79–86, 1951D.

[5] D. Scott, «On optimal and data-based histograms,» *Biometrika,* pp. 605-610, 1979.

[6] R. Fisher, «Iris Data Set,» [Онлайновий]. Available: http://archive.ics.uci.edu/ml/datasets/Iris.

[7] V. Asch, «Macro- and Micro-Averaged Evaluation Measures,» 2012. [Онлайновий]. Available: https://www.clips.uantwerpen.be/~vincent/pdf/microaverage.pdf.

УДК 681.513.8

МЕТОД НЕЧЕТКИХ K-СРЕДНИХ С ОГРАНИЧЕННОЙ МАССОЙ РАБОЧЕЙ ОБЛАСТИ ФОРМИРОВАНИЯ КЛАСТЕРОВ ПРОИЗВОЛЬНОЙ ФОРМЫ

Настенко Е. А., д.б.н., с.н.с., e-mail : nastenko.e@gmail.com
Кафедра биомедицинской кибернетики
"Киевский политехнический институт имени Игоря Сикорского"
Киев, Украина
Уманец В. С., бакалавр, e-mail : 2_bytes@ukr.net
Факультет биомедицинской инженерии
"Киевский политехнический институт имени Игоря Сикорского"
Киев, Украина

Реферат. *Задача определения функциональной связи между биофизическими показателями является составной частью решения актуальной проблемы поиска оптимальных воздействий на биологический объект и не решена в полной мере в настоящее время. Одной из важных задач в этой области является разбиение пространства признаков на области (кластеры), которые относятся к различным функциональным соотношениям, связывающим биофизические показатели, искомые кластеры при этом могут, иметь произвольную форму. Такие кластеры назовем функциональными, в работе ставится задача разработки метода выделения из исходной выборки данных кластеров произвольной формы. Для решения поставленной задачи в работе рассматривается нечеткая версия кластеризации для алгоритма k-средних с ограниченной массой рабочей области формирования кластеров. Оценка количества кластеров проводится по гистограмме частот, для определения оптимального количества столбцов гистограммы обосновывается применение формулы Скотта. Алгоритм позволяет формировать кластеры произвольной конфигурации с получением значения меры принадлежности объекта каждому из кластеров. Эффективность алгоритма продемонстрирована на примере кластеризации набора данных «Ирисы Фишера». Проведено сравнительное тестирование: классический алгоритм k-средних, метод Варда и разработанный алгоритм. Полученные результаты позволили отдать предпочтение в задачах анализа кластеров произвольной формы разработанной в данной работе версии нечеткого k-средних с ограниченной массой рабочей области формирования кластеров. Расчет функции принадлежности позволяет получить дополнительную информацию о структуре кластерных образований, а также осуществить поправки результата кластеризации k-средних с ограниченной массой, что особенно важно для алгоритмов, получающих результат кластеризации за один проход. Относительно сравнения качественных результатов разработанного алгоритма и алгоритма Варда следует отметить, что разработанный алгоритм имеет низкую вычислительную стоимость так как не требует дополнительной памяти для хранения матрицы расстояний и времени на ее перерасчет. Кроме того, разработан алгоритм не имеет проблем, связанных с разрезом дендрограммы для получения кластеров.*

Ключевые слова: *кластеризация, k-средних, мера принадлежности, оценка количества кластеров, нечеткая кластеризация.*

UDC 681.513.8

FUZZY K-MEANS METHOD WITH A LIMITED MASS OF THE WORKING REGION FOR THE FORMATION OF ARBITRARY SHAPED CLUSTERS

Nastenko Ie. A., Doctor of Biological Sciences, Senior Researcher
e-mail: nastenko.e@gmail.com
Department of Biomedical Cybernetics
"Igor Sikorsky Kyiv Polytechnic Institute"
Kiev, Ukraine
Umanets V. S., bachelor
e-mail: 2_bytes@ukr.net
Faculty of Biomedical Engineering
"Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine

Abstract. *The task of determining the functional connection between biophysical indicators is an integral part of the solution of an actual problem of searching for optimal effects on a biological object and has not been fully solved to date. One of the important tasks in this area is the division of the feature space into regions (clusters), which relate to various functional relationships linking biophysical indicators, the desired clusters can have an arbitrary shape. Such clusters will be called functional, and the task is to develop a method for extracting clusters of arbitrary shape from the initial sample. To solve this problem, the paper considers a fuzzy version of clustering for the algorithm of k-means with a limited mass of the working region for clusters' formation. The estimation of number of clusters is carried out according to the histogram of frequencies, to determine the optimum number of columns of the histogram, the application of the Scott formula is justified. The algorithm allows forming clusters of arbitrary configuration and obtaining the value of the object's membership function value for each of the clusters. The efficiency of the algorithm is demonstrated by the example of clustering the Iris Fisher data set. Comparative testing was carried out: classical k-means algorithm, Ward's method and developed algorithm. Obtained results made it possible to give preference to the problems of analyzing clusters of an arbitrary shape developed in this paper, a version of fuzzy k-means with a limited mass of the working region for the formation of clusters. Membership function calculation allows obtaining additional information on the clusters' formation structure, as well as making corrections to the result of clustering of k-means with a limited mass, which is especially important for algorithms that receive the result of clustering in a single pass. Concerning the comparison of the qualitative results of the developed algorithm and the Ward algorithm, it should be noted that developed algorithm has low computational cost since it does not require additional memory to store the distance matrix and time for its recalculation. In addition, developed algorithm has no problems associated with cutting the dendrogram to obtain clusters.*

Keywords: *clustering, k-means, membership function, estimation of number of clusters, fuzzy clustering.*