МАТЕРІАЛИ XXVII МІЖНАРОДНОГО МОЛОДІЖНОГО ФОРУМУ

МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

РАДІОЕЛЕКТРОНІКА ТА МОЛОДЬ У XXI СТОЛІТТІ

2023

ТОМ 7

ХАРКІВ

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
РАДІОЕЛЕКТРОНІКИ

МАТЕРІАЛИ 27-го МІЖНАРОДНОГО МОЛОДІЖНОГО ФОРУМУ

«РАДІОЕЛЕКТРОНІКА І МОЛОДЬ У XXI СТОЛІТТІ»

10-12 травня 2023 р.
том 7

КОНФЕРЕНЦІЯ «КОМП'ЮТЕРНИЙ ЗІР, СИСТЕМНИЙ АНАЛІЗ ТА МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ»

Харків 2023

УДК 004.932+330.46+005.934](06)

27-й Міжнародний молодіжний форум «Радіоелектроніка і молодь у XXI столітті». Зб. матеріалів форуму. Т. 7. Харків: ХНУРЕ. 2023. 243 с.

У збірник включені матеріали 27-го Міжнародного молодіжного форуму «Радіоелектроніка і молодь у XXI столітті».

Видання підготовлено
факультетом інформаційно-аналітичних технологій та менеджменту
Харківського національного університету радіоелектроніки

61166 Україна, Харків, просп. Науки, 14
тел./факс: (057) 7021397

Email: mref21@nure.ua

УДК 004.032.2

GENERAL OVERVIEW OF DATA STREAM CLUSTERING

Помазан В.В.
Науковий керівник – к.т.н., доц. Творошенко І.С.
Харківський національний університет радіоелектроніки, каф. ІНФ,
м. Харків, Україна
тел. +38(068) 367-04-82, e-mail: viktor.pomazan@nure.ua

Data stream clustering is a widely researched problem in data mining and machine learning. This involves grouping data points that arrive sequentially and continuously over time, making it difficult to work with because of the large amounts of data that need to be processed in real-time. This work reviews recent advances in data stream clustering algorithms and techniques, including traditional and online clustering methods, ensemble methods, and stream-based outlier detection, discusses the evaluation metrics for data stream clustering, and highlights some open research challenges in this area.

The task of clustering data flows occupies an important place in modern Data mining, and for its solution today there are many approaches, methods, and algorithms [1–3]. Clustering is related to statistical data processing, as well as to learning tasks without a teacher. Groups of elements are placed in clusters, which are formed based on the distance between them. Clustering can be formulated as a multi-criteria optimization problem. The result of clustering depends on a number of factors: the selected method, data values, the number of formed clusters, and the type of metric used to compare individual elements and clusters with each other [4].

Data stream clustering is the process of grouping incoming data instances, which arrive continuously and potentially in an infinite stream, into clusters in an online and incremental manner.

The main goal of data stream clustering is to extract meaningful patterns and insights from data streams that are too large to be stored in memory or processed offline [2, 3].

During the study of the data stream clustering problem, many researchers and teams have made significant contributions to the development of data stream clustering, including researchers such as S. Guha, R. Rastogi, K. Shim (BIRCH algorithm), M. Garofalakis, J. Gerke, and R. Rastogi (STREAM algorithm), H. Karypis and E.-Kh. Khan (DENSWAVE algorithm), M. Halkidi, Y. Batistakis, M. Vazirgianis (C-Stream algorithm), and many others.

The examination of the most recent investigation and literature concerning this subject revealed that the concerns addressed are not adequately expounded upon, and the primary scientific outcomes have not been put into practical practice and necessitate further investigation.

Data stream clustering algorithms can be classified into two main categories: centroid-based and density-based.

Centroid-based algorithms, such as *k*-means and *k*-medoids, represent each cluster by a central point or prototype and iteratively update these centroids as new data arrives. Density-based algorithms, such as DBSCAN and OPTICS, partition the data space based on the density of data points and identify clusters as regions of high density.

In addition to these traditional clustering algorithms, recent research in data stream clustering has focused on developing novel algorithms that can handle specific challenges of data streams, such as concept drift, noise, and outliers. These include incremental clustering algorithms (CluStream, DenStream), and ensemble clustering algorithms (STREAM, CluELM).

Data stream clustering is a critical research area with numerous applications in various fields. One of the most prominent applications of data stream clustering is in the area of network intrusion detection, where it is used to detect and prevent cyber-attacks. Financial fraud detection is also an important application, where data stream clustering is used to identify patterns of fraudulent behavior in financial transactions [5].

One of the main problems of data stream clustering is concept drift, which occurs when the statistical properties of the data change over time. This can be caused by changes in user behavior, shifts in trends, or changes in the environment. Data stream clustering can also face problems with selecting the optimal window size, cluster size, and a number of clusters. Another issue is accounting for uncertainty and errors that may arise due to incomplete information and inaccuracies in the data stream [3]. Finally, managing the data stream and collecting feedback is necessary to adjust and update clustering models based on the results obtained.

Overall, data stream clustering is an important field of research with a wide range of applications in various domains, including network intrusion detection, sensor data analysis, social media monitoring, and financial fraud detection.

References:

1. Гороховатський, В.О., Гадецька, С.В., Стяглик, Н.І., Власенко, Н.В. (2020). Класифікація зображень на підставі ансамблю статистичних розподілів за класами еталонів для компонентів структурного опису.

2. Гороховатський, В.О., Творошенко, І.С. (2021). Методи інтелектуального аналізу та оброблення даних: навч. посібник.

3. Silva, J.A., Faria, E.R., Barros, R.C., Hruschka, E.R., Carvalho, A.C.D., Gama, J. (2013). Data stream clustering: A survey. *ACM Computing Surveys (CSUR)*, 46(1), 1-31.

4. Yu, S.S., Chu, S.W., Wang, C.M., Chan, Y.K., Chang, T.C. (2018). Two improved k-means algorithms. *Applied Soft Computing*, 68, 747-755.

5. Borlea, I.D., Precup, R.E., Borlea, A.B., Iercan, D. (2021). A unified form of fuzzy C-means and K-means algorithms and its partitional implementation. *Knowledge-Based Systems*, 214, 106731.