Selecting Heterogeneous Features Based on Unified Density-Guided Neighborhood Relation for Complex Biomedical Data Analysis

Lang Zhao[1], Yiqun Zhang[1,4,*], Xiaopeng Luo[2], Yue Zhang[3], Yiu-ming Cheung[4], Kangshun Li[5]
[1]*Guangdong University of Technology,* [2]*BNU-HKBU United International College*
[3]*Guangdong Polytechnic Normal University,* [4]*Hong Kong Baptist University,* [5]*South China Agricultural University*
3120005087@mail2.gdut.edu.cn, yqzhang@gdut.edu.cn, r130201612@mail.uic.edu.cn
zhangyue@gpnu.edu.cn, ymc@comp.hkbu.edu.hk, likangshun@scau.edu.cn

***Abstract*—Biomedical big data are usually high dimensional and collected in the form of a continuous influx of new features. Online Feature Selection (OFS) is a promising way to manage and analyze such data, as OFS circumvents the huge computation cost brought by simultaneously considering all the features, and can also dynamically maintain a distribution-fitting feature subset on the fly. However, almost all the OFS solutions are based on a naive premise that all features are of the same type, overlooking the fact that real biomedical data set usually consists of heterogeneous numerical and categorical features. This paper therefore proposes a new approach to Online Heterogeneous Feature Selection (OHFS), which dynamically maintains a feature subset that maximizes the number of neighborhood sets where all the objects within each neighborhood set are of the same class. To appropriately partition the objects into neighborhood sets, a density-guided relation is proposed, which adaptively forms non-overlapping neighborhood sets by detecting spatially compact objects. A unified density measure is also presented to avoid information loss in processing heterogeneous features. It turns out that the proposed approach features parameter-free, interpretability, and efficiency. It is capable of maintaining a concise feature subset while receiving any type of feature. Extensive experimental evaluations demonstrate its superiority.**

***Index Terms*—Online feature selection, heterogeneous features, distance metric, density measure, supervised learning.**

I. Introduction

Feature selection plays a key role in biomedical data analysis [1]–[3]. The general purpose of feature selection is to search for an optimal feature subset that can appropriately describe the data distribution w.r.t. certain analysis tasks [4]. Online Feature Selection (OFS) becomes a promising way to analyze high-dimensional biomedical data [5], [6], as it can effectively circumvent the curse of dimensionality and incorporate newly generated features [7]. Since categorical features are very common in real data [8], many recent analysis attempts have been attracted [9]–[12]. Table I demonstrates a data set with both categorical and numerical features, where a categorical feature may have multiple qualitative values not embedded in a well-defined distance space [13].

TABLE I: Fragment of a biomedical data set.

ID	Sex (Categorical)	GCS Type (Categorical)	Height (Numerical)	...	Class
1	male	to_speech	190	...	Type8
2	female	to_pain	170	...	Type2
3	male	spontaneoust	175	...	Normal
⋮	⋮	⋮	⋮	⋮	⋮
452	male	none	190	...	Type7

In the literature, an OFS approach [14] has been proposed based on classical rough set theory, which is feasible for the selection of categorical features. The basic idea is that more data objects can be matched with their class labels under the distribution representation of the optimal feature subset. Later, by adopting a similar basic idea but adding an extra feature redundancy analysis procedure via k-greedy search, another OFS approach has also been proposed [15]. They both adopt rough set theory, which is flexible, robust, and convenient to take into account the relationship between the current features and newly arrived ones. Although these two approaches achieve considerable performance in OFS, they are based on the hypothesis that all features are of the same type, which is still not the case for most real biomedical data. To cope with data sets composed of both numerical and categorical features, various Heterogeneous Feature Selection (HFS) approaches have been proposed [16]. Among them, the one proposed in [17] computes lower approximations of classes based on neighborhood rough sets formed through combining Euclidean distances and Hamming distances on numerical and categorical features, respectively, to quantify the significance of heterogeneous features. Later, [18] further introduces entropy to more appropriately distinguish objects in forming neighborhood sets. Differing from them, approaches [19], [20] first discretize numerical features, and then adopt

This work was supported in part by the NSFC under Grants 62102097, 62172112, and 61573157, the NSFC and Research Grants Council (RGC) Joint Research Scheme under Grant N_HKBU214/21, the General Research Fund of RGC under Grants 12201321 and 12202622, the Natural Science Foundation of Guangdong Province under grants 2022A1515011592 and 2023A1515012855, the Science and Technology Program of Guangzhou under grant 202201010548, and by Hong Kong Baptist University under Grant RC-FNRA-IG/18-19/SCI/03. Yiqun Zhang is the corresponding author.

entropy as a measure to select feature subset. As studied by our previous work [21], [22], the awkward information gap between heterogeneous features cannot be well bridged by simply adopting the metric combination and numerical feature discretization, especially in data analysis tasks involving inter-feature relationship. More importantly, all the above HFS approaches are designed for static data only, which are inadequate for Online Heterogeneous Feature Selection (OHFS).

Since the state-of-the-art distance and correlation measures of heterogeneous features [23] are all based on the global data statistics, which will be dynamically updated in OHFS, none of the existing OFS and HFS solutions are directly applicable to the challenging OHFS problem to the best of our knowledge. Motivated by this, we propose an OHFS approach by first defining a more appropriate neighborhood relation guided by the distribution densities of objects, and then unifying the distance metric of categorical and numerical features to form a pertinent density measurement basis for the computation of neighborhood relation. In this approach, instead of forming neighborhood sets for each object, non-overlapping neighborhood sets are formed for the efficient purpose of online learning. To appropriately obtain such neighborhood sets, density and distance are both considered for partitioning the data objects. That is, only locally compact objects with relatively sparse boundaries to the surrounding objects can be partitioned into one neighborhood set.

To form a distance metric that is unified on heterogeneous features, we compute transformation cost between representations of two feature values to indicate their distance. That is, two values of a feature are represented by the conditional probability distributions of another relevant feature. If the transformation cost between these two representations is large, then the two values are considered to be more dissimilar. Since each feature is uniformly represented in the form of a graph, which is considered to be the current most informative form for representing data, to derive the transformation cost, the information of heterogeneous features can be exploited in a homogeneous way to appropriately reflect both the distance and density of data objects for heterogeneous feature selection. Comprehensive experiments including ablation studies, significance tests, and efficiency evaluation have been conducted on various real biomedical data sets to demonstrate the superiority of the proposed approach over six state-of-the-art counterparts. The main contributions of this paper are three-fold:

- To the best of our knowledge, this is the first attempt to tackle the OHFS problem. It relaxes the assumption that streaming features are of the same type. Since it is parameter-free and intuitive, it is promising for biomedical data management and analysis.
- A novel self-adaptive neighborhood relation based on distance and density has been proposed to form compact non-overlapping neighborhood sets. Such a relation is more efficient and effective in boosting OHFS, especially for real data with complex object distributions.
- We unify the distance and density of data objects represented by heterogeneous features by adopting graph-based transformation cost as a measure, which well-bridges the awkward information gap while preserving the intrinsic properties of heterogeneous features.

TABLE II: Frequently used symbols.

Symbols	Explanations
$\mathbf{x}_i \in U$	i-th data object of the universe U;
$\mathbf{f}_r \in F$	r-th feature of feature set F;
$C_m \in C$	m-th class of class set C;
$v_g^r \in V_r$	g-th possible value of $\mathbf{f}_r$'s value domain V_r;
$AN_{F'i}^k$	a set of k objects arranged according to ascending order of their distances to $\mathbf{x}_i$ computed upon F';
$S(F')$	significance of a feature set F';
$\mathcal{F}$	a set of mixed numerical and categorical features;
$\mathcal{D}(\cdot,\cdot;\mathcal{F}')$	distance between two data objects computed upon $\mathcal{F}'$;
$\mathcal{D}^r(\cdot,\cdot)$	distance between two values of $\mathbf{f}_r$;
μ_i	density gap of $\mathbf{x}_i$ computed upon object densities;
ρ_i	density of $\mathbf{x}_i$;
$R_{\mathcal{F}'}^{\mu}(\mathbf{x}_i)$	neighborhood set of $\mathbf{x}_i$ guided by density gap μ_i;

The rest of this paper is organized as follows. Section II provides preliminaries, while Section III presents the proposed approach in detail. Experimental results are demonstrated in Section IV with in-depth analysis. Finally, Section V gives the concluding remarks.

II. Preliminaries

This section presents necessary pre-definitions of OHFS, including common notations and brief formulation of the neighborhood relation-based feature selection in Section II-A and online feature selection in Section II-B. A data set for feature selection can be viewed as an information system $IS = (U, F, C, V)$, where the universe $U = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ is a set of n data objects, $F = \{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_d\}$ is a set of d features, $C = \{C_1, C_2, ..., C_l\}$ is a set of l classes with C_m containing all the objects belonging to m-th class, and $V = \{V_1, V_2, ..., V_d\}$ stores the value domain corresponding to each feature with V_t stores possible values of t-th feature. Table II sorts out frequently used symbols in this paper.

A. Neighborhood Relation for Feature Selection

Neighborhood relation is commonly adopted for categorical or heterogeneous feature selection. Before the feature selection, we let each object $\mathbf{x}_i$ find a neighborhood set $R_{F'}(\mathbf{x}_i)$ composed of objects that are more closer to $\mathbf{x}_i$, where the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ reflected by feature subset $F' \in F$ is denoted as $D(\mathbf{x}_i, \mathbf{x}_j; F')$. Two conventional neighborhood relations are the k-nearest and δ-radius relations, i.e., $R_{F'}^k(\mathbf{x}_i) = AN_{F'i}^k$ and $R_{F'}^{\delta}(\mathbf{x}_i) = \{\mathbf{x}_j \mid D(\mathbf{x}_i, \mathbf{x}_j; F') \leq \delta\}$, respectively, where $j \in \{1, 2, ..., n\}$, and

$$AN_{F'i}^k = \{\mathbf{x}_{\langle i,0\rangle}, \mathbf{x}_{\langle i,1\rangle}, ..., \mathbf{x}_{\langle i,k-1\rangle}\} \tag{1}$$

is the collection of the top k objects that are more similar to $\mathbf{x}_i$ among all n data objects. The subscript $\langle i, h\rangle$ indicates that the corresponding object ranks hth in the degree of similarity to $\mathbf{x}_i$. Therefore, the objects in $AN_{F'i}^k$ are actually arranged in ascending distance order, i.e., $D(\mathbf{x}_i, \mathbf{x}_{\langle i,0\rangle}; F') < D(\mathbf{x}_i, \mathbf{x}_{\langle i,1\rangle}; F') < ... < D(\mathbf{x}_i, \mathbf{x}_{\langle i,k-1\rangle}; F')$. Note that we let

$\mathbf{x}_{\langle i,0\rangle} \equiv \mathbf{x}_i$. For simplicity without causing ambiguity, we use $R_{F'}(\mathbf{x}_i)$ to generally indicate neighborhood relation.

The above neighborhood relations act to reflect the quality of feature subset F' w.r.t. class labels, and then high-quality features can be selected accordingly. To use the neighborhood set for feature selection, the lower and upper approximation of data objects can be computed according to neighborhood relation $R_{F'}(\mathbf{x}_i)$ w.r.t. each class of objects C_m by

$$\underline{U}_{F'm} = \{\mathbf{x}_i \mid R_{F'}(\mathbf{x}_i) \subseteq C_m\} \tag{2}$$

$$\overline{U}_{F'm} = \{\mathbf{x}_i \mid R_{F'}(\mathbf{x}_i) \cap C_m \neq \emptyset\} \tag{3}$$

respectively, where $i \in \{1, 2, ..., n\}$ and $m \in \{1, 2, ..., l\}$. Based on Eqs. (2) and (3), the overall data set U can be partitioned into positive, boundary, and negative regions w.r.t. class C_m, which can be represented as $U^+_{F'm} = \underline{U}_{F'm}$, $U^{\sim}_{F'm} = \overline{U}_{F'm} - \underline{U}_{F'm}$, and $U^-_{F'm} = U - \overline{U}_{F'm}$, respectively. Given an F', the corresponding overall positive region is thus

$$U^+_{F'} = \bigcup_{m=1}^{l} U^+_{F'm} = \bigcup_{m=1}^{l} \underline{U}_{F'm}. \tag{4}$$

It is intuitive that a preferred F' tends to maximize the size of $U^+_{F'}$, which indicates that F' can maximize the certainty of objects w.r.t. all the class labels.

B. Online Feature Selection

For OFS, it is commonly assumed that the features flow in one by one at each time-stamp t while the number of data objects n remains fixed. It is also worth noting that a common setting is to make the discarded features never come back. During the OFS process, Eq. (4) is commonly used to indicate the significance of a feature subset by

$$S(F') = \frac{|U^+_{F'}|}{n} \tag{5}$$

where $|\cdot|$ obtains the cardinality of a set, and thus $|U^+_{F'}|$ reflects the number of objects in the positive regions. Eq. (5) computes the proportion of objects that are partitioned into the positive regions based on F'. Accordingly, features can be divided into three types by the following definitions for guiding OFS.

Definition 1. *Significant feature: Given C, F', and a new feature $\mathbf{f}_t$ at time-stamp t, if $S(F' \cup \mathbf{f}_t) - S(F') > 0$, then $\mathbf{f}_t$ is a significant feature for F'.*

Definition 2. *Redundant feature: Given C, F', and a new feature $\mathbf{f}_t$ at time-stamp t, if $S(F' \cup \mathbf{f}_t) - S(F') = 0$, then $\mathbf{f}_t$ is a redundant feature for F'.*

Definition 3. *Irrelevant feature: Given C, F', and a new feature $\mathbf{f}_t$ at time-stamp t, if $S(F' \cup \mathbf{f}_t) - S(F') < 0$, then $\mathbf{f}_t$ is an irrelevant feature for F'.*

Based on the above definitions, a significant $\mathbf{f}_t$ should be incorporated by F', an irrelevant $\mathbf{f}_t$ should be rejected by F', and redundant features in F' should be discarded, to maintain a maximum $S(F')$ at each time-stamp t during the OFS. Note that we will use $\mathcal{F}$ instead of F in Section III to indicate a heterogeneous feature set.

III. Propose Method

This section first formulates the OHFS problem, then presents the proposed self-adaptive neighborhood relation, unified distance metric, and the whole feature selection algorithm in Section III-A, III-B, and III-C, respectively.

For a heterogeneous feature set $\mathcal{F}$, we have $d = d_c + d_n$ where d_n and d_c are the number of numerical and categorical features, respectively. Value domain of a categorical feature $\mathbf{f}_r$ is denoted as a set of unique values $V_r = \{v^r_1, v^r_2, ..., v^r_{\mathsf{v}^r}\}$ where v^r is the number of $\mathbf{f}_r$'s possible values. Based on Eq. (5), objective of OHFS can be formalized as maintaining $\mathcal{F}'_t$ with significant features at each time-stamp t to maximize the significance $S(\mathcal{F}'_t)$, which can be written as

$$\mathcal{F}'_t = \arg\max_{\mathcal{F}'} S(\mathcal{F}') \quad \text{s.t.} \quad \mathcal{F}' \subseteq \mathcal{F}'_{t-1} \cup \mathbf{f}_t. \tag{6}$$

According to Section II-B, the positive region $U^+_{\mathcal{F}'}$ for calculating $S(\mathcal{F}'_t)$ is based on $R_{\mathcal{F}'}$, and obtaining $R_{\mathcal{F}'}(\mathbf{x}_i)$ requires the inter-object distances, we thus adopt a common form of inter-object distance w.r.t. $\mathcal{F}'$ as

$$\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j; \mathcal{F}') = \sqrt{\sum_{\mathbf{f}_r \in \mathcal{F}'} \mathcal{D}^r(x^r_i, x^r_j)^2}. \tag{7}$$

where $x^r_i \in V_r$ is the value of $\mathbf{x}_i$ on $\mathbf{f}_r$, and $\mathcal{D}^r\left(x^r_i, x^r_j\right)$ measures the distance between two feature values x^r_i and x^r_j of $\mathbf{f}_r$. Then we introduce how to define $R_{\mathcal{F}'}(\mathbf{x}_i)$ and $\mathcal{D}^r(x^r_i, x^r_j)$ in the following two sub-sections.

A. SANR: Self-Adaptive Neighborhood Relation

Existing neighborhood relations described in Section II-A form n neighborhood sets that may overlap each other, which may cause high computation cost for considering each current feature in OHFS. Moreover, for unevenly distributed data objects, the k-nearest relation may incorporate very dissimilar objects into the same neighborhood set, while the δ-radius relation may form neighborhood sets crossing class boundaries if the value of δ is set too large. Inappropriate neighborhood sets surely influence the performance of the corresponding feature selection.

To appropriately distinguish class boundaries and save computation cost, we propose a new neighborhood relation based on density to adaptively form non-overlapped compact neighborhood sets that may have different sizes. To obtain such sets, we first select representative data objects for the corresponding local regions based on *density gap*.

Definition 4. *Density gap: Density gap μ_i of object $\mathbf{x}_i$ with density ρ_i is the minimum distance between $\mathbf{x}_i$ and another object $\mathbf{x}_j$ with higher density ρ_j, which can be written as*

$$\mu_i = \min \mathcal{D}(\mathbf{x}_i, \mathbf{x}_j; \mathcal{F}') \tag{8}$$

$$\textit{s.t.} \quad \mathbf{x}_j \in U \backslash \mathbf{x}_i, \quad \rho_j > \rho_i, \quad \textit{and} \quad \rho_i = \frac{k_i}{\mathcal{D}(\mathbf{x}_i, \mathbf{x}_{\langle i,k_i\rangle}; \mathcal{F}')}.$$

In Definition 4, the notation $U \backslash \mathbf{x}_i$ indicates the universe excluding $\mathbf{x}_i$, density ρ_i is computed as the number of neighbor objects (i.e., k_i) per the radius $\mathcal{D}(\mathbf{x}_i, \mathbf{x}_{\langle i,k_i\rangle}; \mathcal{F}')$ by treating

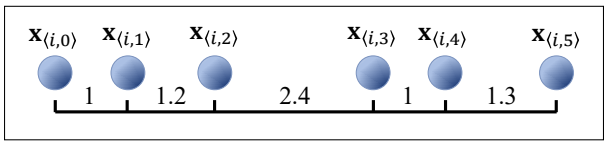

Fig. 1: A toy example with $n = 6$ to explain the determination of k_i. Please note that $\mathbf{x}_{\langle i,0\rangle} \equiv \mathbf{x}_i$, $AN^6_{\mathcal{F}'i} = \{\mathbf{x}_{\langle i,0\rangle}, \mathbf{x}_{\langle i,1\rangle}, \mathbf{x}_{\langle i,2\rangle}, \mathbf{x}_{\langle i,3\rangle}, \mathbf{x}_{\langle i,4\rangle}, \mathbf{x}_{\langle i,5\rangle}\}$, each sphere represents a data object, and the scaled line indicates distances between every pair of adjacent data objects. By going through $AN^6_{\mathcal{F}'i}$ from the left to the right, it can be found that $\mathbf{x}_{\langle i,4\rangle}$ is the first object satisfying $\mathcal{D}(\mathbf{x}_i, \mathbf{x}_{\langle i,4\rangle}; \mathcal{F}')/4 < \mathcal{D}(\mathbf{x}_i, \mathbf{x}_{\langle i,4-1\rangle}; \mathcal{F}')/(4-1)$, thus $q = 4$ and $k_i = q - 2 = 2$.

$\mathbf{x}_i$ as the circle center, where $\mathbf{x}_{\langle i,k_i\rangle}$ is an object ranks k_ith in terms of closeness to $\mathbf{x}_i$ among all the n data objects, as defined in Eq. (1). For the data object with the highest density, we set its density gap as $\max \mathcal{D}(\mathbf{x}_i, \mathbf{x}_j; \mathcal{F}')$ with $\mathbf{x}_j \in U$.

As shown in Fig. 1, a mechanism is proposed to adaptively determine k_i for different $\mathbf{x}_i$ to ensure that only significantly close objects are considered for the computation of density. Specifically, we first obtain the distance ascending neighbor set $AN^n_{\mathcal{F}'i}$ for $\mathbf{x}_i$ by considering all the n objects in the universe U according to Eq. (1). Then by going through $AN^n_{\mathcal{F}'i}$ from the left to the right, q is determined by finding the object $\mathbf{x}_{\langle i,q\rangle}$ that first satisfies

$$\frac{\mathcal{D}(\mathbf{x}_i, \mathbf{x}_{\langle i,q\rangle}; \mathcal{F}')}{q} < \frac{\mathcal{D}(\mathbf{x}_i, \mathbf{x}_{\langle i,q-1\rangle}; \mathcal{F}')}{q-1}, \tag{9}$$

then we determine $k_i = q - 2$, because $\mathbf{x}_{\langle i,q-2\rangle}$ is the last object in $AN^n_{\mathcal{F}'i}$ that is significantly close to $\mathbf{x}_{\langle i,0\rangle}$.

It is intuitive that an ideal representative object of a neighborhood set should be surrounded by its neighbors with lower density, and should also have a relative long distance to the other representative objects with higher density. According to Definition 4, an object with a higher density gap is more suitable to become a representative object. Therefore, we first rank objects in descending order of their density gaps, and then form neighborhood sets for the objects in turn by

$$R^{\mu}_{\mathcal{F}'}(\mathbf{x}_i) = AN^{k_i}_{\mathcal{F}'i} \setminus \left\{ \bigcup_{\mu_g > \mu_i} AN^{k_g}_{\mathcal{F}'g} \right\} \tag{10}$$

until all data objects are incorporated by the neighborhood sets. The computation process of significance $S(\mathcal{F}')$ based on the proposed SANR is summarized in Algorithm 1.

As $R^{\mu}_{\mathcal{F}'}(\mathbf{x}_i)$ relies on the object-level distance $\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j; \mathcal{F}')$ (see Eq. (7)), how to appropriately define $\mathcal{D}^r(x^r_i, x^r_j)$ on heterogeneous features $\mathcal{F}'$ to form $\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j; \mathcal{F}')$ is another key problem that will be solved in the next subsection.

B. UDM: Unified Distance Metric

To unify the distance $\mathcal{D}^r(x^r_i, x^r_j)$ w.r.t. different heterogeneous features $\mathbf{f}_r$, we adopt transformation cost that quantifies the cost for converting a Conditional Probability Distribution (CPD) into another as a measure. We first specify the CPD, and then formulate the distance. Finally, we show that the distance is unified for categorical and numerical features.

Algorithm 1 Significance of feature subset based on SANR.

Input: U, C, V, $\mathcal{F}'$, and $\mathcal{D}^r$.
Output: $S(\mathcal{F}')$.
1: **for** $i = 1 \to n$ **do**
2: Compute μ_i for $\mathbf{x}_i$ according to Eqs. (8) and (9);
3: **end for**
4: **for** μ_is in descending order **do**
5: Obtain $R^{\mu}_{\mathcal{F}'}(\mathbf{x}_i) \notin \emptyset$ according to Eq. (10);
6: **end for**
7: Compute $S(\mathcal{F}')$ according to Eqs. (2), (4), and (5);

CPD of a feature $\mathbf{f}_s$ given a possible value v^r_g of another feature $\mathbf{f}_r$ can be written as

$$\mathbf{p}^{r\leftarrow s}_g = [p(v^s_1|v^r_g), p(v^s_2|v^r_g), ..., p(v^s_{\mathsf{v}^s}|v^r_g)]^{\top} \tag{11}$$

where $p(v^s_h|v^r_g)$ is conditional probability of v^s_h as given v^r_g. Superscript $r \leftarrow s$ and subscript g indicate that such a CPD is utilized to represent the g-th possible value of $\mathbf{f}_r$ by the values of $\mathbf{f}_s$. For simplicity, we denote $r \leftarrow s$ as rs hereinafter.

As studied by most categorical data distance measures, the difference between two such CPDs, e.g., $\mathbf{p}^{rs}_g$ and $\mathbf{p}^{rs}_q$ can effectively reflect the dissimilarity between two possible values, i.e., v^r_g and v^r_q. Thus the Earth Mover's Distance (EMD) [24], which has been proposed for computing the transformation cost between two histogram descriptors of images, is adopted to quantify the difference between $\mathbf{p}^{rs}_g$ and $\mathbf{p}^{rs}_q$ according to the graph structure of $\mathbf{f}_s$ that can be constructed according to our previous work [23]. That is, a graph structure of $\mathbf{f}_s$ is with all the v^s possible values (as nodes) connected by $\mathsf{v}^s(\mathsf{v}^s - 1)/2$ edges with identical length "1". The edge lengths indicate the transformation cost per the quantity that needs to be transformed between nodes. The necessity for introducing graph representation is to unify the transformation cost for both numerical and categorical features, which will be discussed later in Theorem 2.

Accordingly, distance between v^r_g and v^r_q reflected by $\mathbf{f}_s$ can be expressed according to the graph structure of $\mathbf{f}_s$ below[1]:

$$\mathcal{D}^{rs}(v^r_g, v^r_q) = \max((\mathbf{p}^{rs}_g - \mathbf{p}^{rs}_q), \mathbf{0}) \cdot \mathbf{1} \tag{12}$$

where $\max(\cdot,\cdot)$ compares each pair of corresponding bits of two vectors and reserves the maximum value (i.e., quantity that will be transformed), and $\mathbf{1}$ is a v^s-dimensional vector with all its values equal to 1 (i.e., cost per the quantity of each bit that need to be transformed). Due to different degrees of inter-feature dependence, different features $\mathbf{f}_s$ may contribute differently according to their importance w^{rs} in forming the overall distance $\mathcal{D}^r(v^r_g, v^r_q)$ between v^r_g and v^r_q by

$$\mathcal{D}^r(v^r_g, v^r_q) = \sum_{\mathbf{f}_s \in \mathcal{F}'} \mathcal{D}^{rs}(v^r_g, v^r_q) \cdot w^{rs}. \tag{13}$$

[1]Due to space limitation, please refer to [23] for more derivation details.

Thus we further extend Eq. (12) to measure the inter-feature dependence as the importance w^{rs}, which can be written as

$$w^{rs} = \frac{\sum_{g=1}^{\mathsf{v}^r-1}\sum_{q=g+1}^{\mathsf{v}^r}\mathcal{D}^{rs}(v_g^r, v_q^r)}{\mathsf{v}^r\left(\mathsf{v}^r-1\right)/2}. \tag{14}$$

According to [21], possible values of a categorical feature can be viewed as different concepts. Eq. (14) actually quantifies average inter-concept distances of $\mathbf{f}_r$ reflected by $\mathbf{f}_s$. We take an extreme example to explain Eq. (14). When $\mathbf{f}_r$ and $\mathbf{f}_s$ are identical, they are perfectly interdependent, and the corresponding $\mathcal{D}^{rs}(v_g^r, v_q^r)$ always reaches the maximum "1" for arbitrary combinations of g and q with $g \neq q$ according to Eq. (12). Therefore, the corresponding w^{rs} also reaches the maximum "1" indicating a 100% inter-feature dependence.

Based on Eqs. (12) - (14), object-level distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ on $\mathcal{F}'$ can be computed through Eq. (7). We prove that the defined distance is a unified distance metric.

Theorem 1. *$\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j; \mathcal{F}')$ is a distance metric.*

Proof. It is clear that Eq. (12) is a metric, and thus Eq. (13) based on Eq. (12) is also a metric. Since Eq. (7) is computed by performing finite arithmetic operations based on Eq. (13), $\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j; \mathcal{F}')$ satisfies all the metric properties:
(1) $\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j; \mathcal{F}') \geq 0$; $\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j; \mathcal{F}') = 0$ iff $\mathbf{x}_i = \mathbf{x}_j$;
(2) $\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j; \mathcal{F}') = \mathcal{D}(\mathbf{x}_j, \mathbf{x}_i; \mathcal{F}')$;
(3) $\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j; \mathcal{F}') \leq \mathcal{D}(\mathbf{x}_i, \mathbf{x}_l; \mathcal{F}') + \mathcal{D}(\mathbf{x}_l, \mathbf{x}_j; \mathcal{F}')$; □

Theorem 2. *The proposed distance metric treats categorical and numerical features in a unified way w.r.t. priori knowledge that distance space of each numerical feature is an independent one-dimensional continuous*[2] *Euclidean distance space with value domain* $[0, 1]$.

Proof. Distance space of a numerical feature $\mathbf{f}_r$ can be viewed as a graph with $\mathsf{v}^r \to \infty$ linearly arranged nodes connected by $\mathsf{v}^r - 1$ edges with identical length. Due to the independence, we have $w^{rs} \equiv 0$ with $r \neq s$, and thus we should only consider the distance under the case $r = s$, where $w^{rs} = \lim_{\mathsf{v}^r\to\infty}(\sum_{g=1}^{\mathsf{v}^r-1} 1/(\mathsf{v}^r - 1))/1 = 1$ according to Eq. (14). Here, $\mathsf{v}^r \to \infty$ is the number of possible values, and the denominator "1" indicates that there is only one pair of concepts for a numerical feature, i.e., "0" for "none" and "1" for "yes" [21]. Accordingly, Eq. (13) degrades to $\mathcal{D}^r(v_g^r, v_q^r) = \mathcal{D}^{rr}(v_g^r, v_q^r) = |v_g^r - v_q^r|$ according to the graph structure, which is equivalent to Euclidean distance. □

C. Overall Feature Selection Algorithm

We then introduce how to analyze the significance and redundancy of heterogeneous features during OHFS. Given an arbitrary feature subset $\mathcal{F}'_{t-1}$, now we consider a new feature $\mathbf{f}_t$ using the Algorithm 2 named USO, which first performs significance analysis to $\mathbf{f}_t$ based on $\mathcal{F}'_{t-1}$. If $\mathbf{f}_t$ is significant, it will be incorporated to form $\mathcal{F}'_t$. If $\mathbf{f}_t$ is an irrelevant feature, it will be simply passed. If $\mathbf{f}_t$ is redundant, redundancy analysis will be further conducted and may yield the removal of any features in $\mathcal{F}'_{t-1} \cup \mathbf{f}_t$ that are no longer significant upon the arrival of $\mathbf{f}_t$ and feature removals.

[2]Numerical integer features are treated as categorical ordinal features here.

Algorithm 2 USO: UDM and SANR-based OHFS.

Input: U, C, V, $\mathcal{D}^r$ defined by Eq. (13), $\forall \mathcal{F}'_{t-1}$, $\mathcal{F} = \mathcal{F}\backslash\mathcal{F}'_{t-1}$, and $S(\mathcal{F}'_{t-1})$ computed by Algorithm 1.
Output: Optimal $\mathcal{F}'_t$ at the current time-stamp t.

```
while F ≠ ∅ do
    Fetch f_t ∈ F, compute S(F'_{t-1} ∪ f_t) by Algorithm 1;
    if S(F'_{t-1} ∪ f_t) > S(F'_{t-1}) then
        F'_t ← F'_{t-1} ∪ f_t and S(F'_t) ← S(F'_{t-1} ∪ f_t);
    end if
    if S(F'_{t-1} ∪ f_t) = S(F'_{t-1}) then
        F'_t ← F'_{t-1} ∪ f_t;
        for each f ∈ F'_t do
            Compute S(F'_t\f) by Algorithm 1;
            if S(F'_t\f) ≥ S(F'_t) then
                S(F'_t) ← S(F'_t\f) and F'_t ← F'_t\f;
            end if
        end for
    end if
    F ← F\f_t and t ← t + 1;
end while
```

Theorem 3. *Time complexity of USO is $O(d_c^3 n + d_c^2 n^2 + d_c n^2 \log n)$ at any time-stamp t.*

Proof. For worst-case analysis, we assume that all the features are categorical, i.e., $d_c = |\mathcal{F}'_t|$, and $\mathsf{V} = \max(\mathsf{v}^1, \mathsf{v}^2, ..., \mathsf{v}^{d_c})$. Distance matrices $M_t = \{\mathbf{M}_1, \mathbf{M}_2, ..., \mathbf{M}_{d_c}\}$ corresponding to each feature in $\mathcal{F}'_t$ should be updated in advance before each time we compute significance S in USO Algorithm, and significance will be computed at most d_c times. Then we analyze the complexity of updating M_t and S once.

The time complexity for updating M_t is $O(nd_c^2 + \mathsf{V}^3 d_c^2)$. Please refer to Theorem 2 in [23] for detailed analysis.

To compute S, we should first compute a $n \times n$ distance matrix for the objects and sort all the rows in the matrix in ascending order with $O(n^2 d_c + n^2 \log n)$. Then for each of the n objects, we should consider at most the remainder $n-1$ ones to determine its k_i value for obtaining the density ρ_i, and should scan the remainder $n-1$ objects to find the one with higher density and the shortest distance to it to obtain the density gap, which take complexity $O(n^2)$. The n density gaps should be sorted in descending order with $O(n \log n)$. Then, at most n neighborhood sets can be formed in order, and each neighborhood set will consider at mots k_i objects in the already sorted distances, with complexity $O(nk_i)$. Overall complexity for computing S is thus $O(n^2 d_c + n^2 \log n + n^2 + n \log n + nk_i)$, which can be simplified to $O(n^2 d_c + n^2 \log n)$.

Since M_t and S will be computed at most d_c times, the overall time complexity at time-stamp t is $O(nd_c^3 + \mathsf{V}^3 d_c^3 + n^2 d_c^2 + d_c n^2 \log n)$. As V is a very small constant, the final time complexity is $O(d_c^3 n + d_c^2 n^2 + d_c n^2 \log n)$. □

Note that during the OHFS, the size of $\mathcal{F}'_t$ will become

TABLE III: Statistics of experimental data sets.

Data sets	Abbr.	d_n	d_c	d	n	Source
SCADI	SC	1	205	206	70	[30]
ARrhythmia	AR	206	73	279	452	[30]
DARWIN	DA	425	25	450	174	[30]
MUsk	MU	168	0	168	476	[30]
Period **C**hanger	PC	1177	0	1177	90	[30]
TOxicity	TO	1200	0	1200	171	[30]
MADELON	MA	0	500	500	600	[31]
HIVA	HI	0	1617	1617	384	[32]

stable quickly after a certain number of time-stamps [25]. That is, the value of d_c in Theorem 3 can be viewed as a small constant (see Table VI), which indicates that USO does not bring much extra computation cost than the existing efficient counterparts [26], [27] with time complexity $O(d_c n^2 \log n)$.

IV. Experimental Results

A. Experimental Settings

Three types of experiments: (1) comparative study, (2) ablation study, and (3) efficiency evaluation, are designed to comprehensively evaluate the efficacy of the proposed USO.

Seven counterparts including one conventional approach, i.e., OS-NRRASR-SA (ONS) [14], and six state-of-the-art approaches, i.e., MRMS [15], A3M [26], OFS-Density (O-Dense) [27], OSFI [28], OSFS-ET (OET) [25] and GapKnn [29], are chosen for comparison. Hyper-parameters of the counterparts (if any) are set according to the corresponding source papers. For the counterparts infeasible in handling heterogeneous features, we let them process numerical and categorical features using Euclidean and Hamming distances, respectively [17], to form two types of neighborhood sets, which are then combined for feature selection.

Eight data sets including three mixed, three pure numerical, and two pure categorical data sets are utilized for the experiments. Statistics of the data sets are shown in Table III. During the experiment, one feature is randomly fetched as the streaming feature without replacement at the current time-stamp, until all features are exhausted.

Four evaluation metrics are utilized to quantify the feature selection performance. We feed the selected feature subset at the final time-stamp to two conventional classifiers, i.e., Support Vector Machine (SVM) [33] and K-Nearest Neighbor (KNN) [34], and then use classification accuracy to form two evaluation metrics, i.e., Acc@SVM and Acc@KNN. Ten-fold cross-validation is used and average accuracy is reported. Bonferroni-Dunn (BD) test with Critical Difference (CD) intervals is chosen to visualize the significance test results. Accuracy per Feature (ApF) is also computed to comprehensively reflect the effectiveness and efficiency of a feature selection approach. All the experiments are coded with Python 3.9.

B. Comparative Study

Performance in terms of Acc@SVM and Acc@KNN are shown in Tables IV and V, respectively. The best and the second-best results are highlighted using boldface and underline, respectively. The "$\overline{\text{Acc}}$" and "$\overline{\text{Rank}}$" rows report the average accuracy and average rank on all the data sets. The corresponding BD significance tests are visualized based on CD interval in Fig. 2. We also report the average number of selected features during the OFS in Table VI with the "Ave" row reporting the average of each column in the table.

TABLE IV: Comparison of Acc@SVM performance.

Data	USO	ONS	MRMS	A3M	O-Dense	OSFI	OET	GapKnn
SC	**0.8285**	0.5857	0.6000	0.5857	0.4428	0.3714	0.7000	0.6428
AR	**0.6667**	0.5690	0.5976	0.5976	0.6071	0.5642	0.6380	0.5952
DA	**0.9001**	0.1879	0.0895	0.8882	0.7133	0.4852	0.7238	0.8356
MU	0.6019	0.4259	0.4378	0.5404	0.5255	0.4304	**0.6077**	0.5361
PC	0.6888	**0.7000**	0.6666	0.6888	0.6555	0.6555	0.6888	0.6888
TO	**0.6705**	**0.6705**	0.6235	0.5764	0.6588	**0.6705**	0.6647	0.6307
MA	0.5235	0.4700	0.4916	0.5233	**0.5300**	0.4783	0.5116	0.5133
HI	**0.9634**	**0.9634**	0.9610	**0.9634**	**0.9634**	**0.9634**	**0.9634**	0.9608
$\overline{\text{Acc}}$	**0.7304**	0.5715	0.5585	0.6704	0.6371	0.5774	0.6873	0.6754
$\overline{\text{Rank}}$	**1.3750**	4.7500	6.0000	3.5000	4.2500	5.6250	2.6250	4.5000

TABLE V: Comparison of Acc@KNN performance.

Data	USO	ONS	MRMS	A3M	O-Dense	OSFI	OET	GapKnn
SC	**0.8143**	0.5714	0.5571	0.4571	0.4142	0.2857	0.5857	0.5285
AR	**0.6857**	0.5214	0.5309	0.5571	0.5666	0.3595	0.6142	0.5571
DA	**0.8941**	0.5656	0.6271	0.8588	0.6908	0.4732	0.7218	0.7941
MU	**0.6241**	0.5503	0.5197	0.5065	0.5647	0.5460	0.6195	0.5000
PC	0.6222	0.6111	0.6111	0.5666	0.5999	**0.6666**	0.6111	0.5333
TO	0.5660	0.5545	0.5081	0.4839	0.5310	**0.6117**	0.4385	0.4915
MA	**0.5300**	0.4850	0.5016	**0.5300**	**0.5300**	0.4783	0.5116	0.5133
HI	**0.9634**	**0.9634**	0.9610	0.9476	**0.9634**	**0.9634**	**0.9634**	0.9608
$\overline{\text{Acc}}$	**0.7125**	0.6028	0.6021	0.6135	0.6076	0.5481	0.6332	0.6098
$\overline{\text{Rank}}$	**1.2500**	4.3750	5.2500	5.2500	3.7500	5.0000	3.3750	5.6250

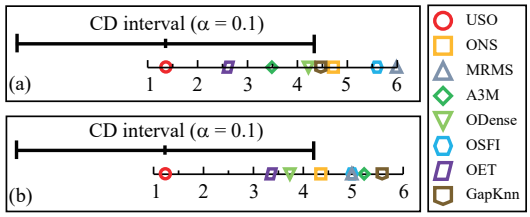

Fig. 2: BD test based on the results in (a) Table IV and (b) Table V. Approaches ranked outside the CD intervals are believed to perform significantly differently from USO.

It can be observed from Tables IV and V that USO outperforms the other counterparts in general, which illustrates its effectiveness. More specifically, USO outperforms the other counterparts on all the mixed data sets (i.e., SC, AR, and DA) indicating that USO can successively fuse the information provided by heterogeneous features as the fundamental distance metric is unified on numerical and categorical features. For the remainder five pure data sets, performance of USO always ranks in top-3, which is not as superior as on mixed data sets, but still very competitive as the significance of feature subset is computed more appropriately based on SANR proposed in Section III-A. That is, during the forming of neighborhood set, density gap acts to finely detect boundary regions among compact object clusters, which are also probably decision boundaries among classes. To sum up, USO is competent in handling any type of features.

We also conducted BD tests according to [35] to the results

TABLE VI: Comparison of the number of selected features.

Data	USO	ONS	MRMS	A3M	O-Dense	OSFI	OET	GapKnn
SC	9.1	7.1	6.5	5.4	5.6	1.0	2.6	5.0
AR	9.8	2.0	1.9	7.6	9.6	3.5	16.7	7.7
DA	8.3	1.0	1.0	15.5	10.0	1.0	31.5	15.4
MU	16.1	3.0	2.5	11.5	13.6	8.4	19.5	11.5
PC	1.0	2.6	2.6	9.4	7.5	1.0	5.8	9.0
TO	1.0	2.5	2.4	9.2	6.6	1.0	13.3	10.2
MA	2.3	3.0	3.4	11.0	10.0	3.0	1.8	11.1
HI	10.0	10.0	10.0	16.0	19.0	2.0	2.0	3.0
Ave	7.2	3.9	3.8	10.7	10.2	2.6	11.7	9.1

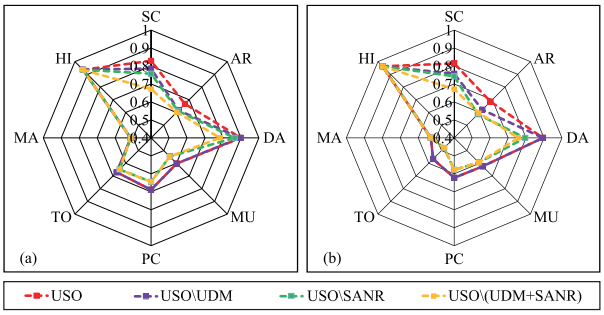

Fig. 3: Performance of different ablated versions of USO evaluated by (a) Acc@SVM and (b) Acc@KNN.

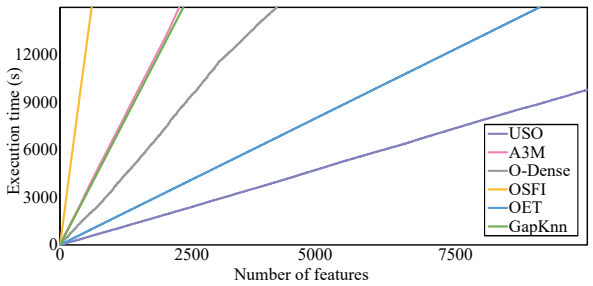

Fig. 4: Comparison of execution time@number of features (d).

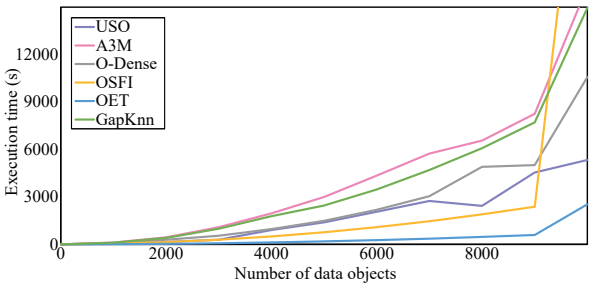

Fig. 5: Comparison of execution time@number of objects (n).

in "$\overline{\text{Rank}}$" rows of Tables IV and V, and visualize the test results based on CD interval in Fig. 2. The length of the CD interval is 3 for comparing eight approaches on eight data sets with $\alpha = 0.1$. The compared approaches are considered to perform significantly differently from USO if they ranked outside the CD interval. It can be seen that USO performs significantly better than most of the counterparts, although most of them are recent state-of-the-arts.

It is worth noting that, although USO does not always significantly outperform the state-of-the-art A3M, OET, and O-Dense, they rely on a larger number of features (see Table VI) to achieve a competitive accuracy. We compute Accuracy per Feature (ApF) by ApF $= \overline{\text{Acc.}}/\text{Ave}$ to more comprehensively compare their performance where "./" means to divide the values of the corresponding positions of two vectors. ApF of USO w.r.t. Acc@SVM is 0.7304/7.2=0.1014, which is obviously higher than that of A3M (0.0626), OET (0.0587), and GapKnn (0.0742). This indicates that USO achieves superior accuracy with a more concise feature subset.

C. Ablation Study

Four ablated versions of USO are compared to verify the effectiveness of its core components. UDM and SANR of USO are replaced with the conventional combination of Euclidean and Hamming distance [17] and δ-radius neighborhood relation to form USO\UDM and USO\SANR, respectively. Both UDM and SANR are replaced to form USO\(UDM+SANR).

It can be observed from Fig. 3 that the accuracy of USO, USO\UDM, USO\SANR, and USO\(UDM+SANR) increases in order on all the mixed data sets (i.e., SC, AR, DA), which intuitively illustrates the effectiveness of UDM and SANR. For pure numerical and categorical data sets, the performance is the same for USO and USO\UDM, which proves that UDM effectively unifies the numerical and categorical distances. Moreover, due to the sparsity of categorial valued objects, UDM and SANR do not act in forming the neighborhood sets, and thus the feature selection results of the four USO versions are the same on all the categorical data sets, i.e., MA and HI.

D. Efficiency Evaluation

To validate the efficiency of USO, its execution time is compared with five state-of-the-art counterparts including A3M, O-Dense, OSFI, OET, and GapKnn, on two large synthetic mixed data sets with increasing d and n. For Fig. 4, n is fixed at 300, and a categorical feature with a number of possible values $\mathsf{v}^t \in \{3, 4, 5, 6\}$ or a numerical feature is randomly generated with equal probability to be the new streaming feature $\mathbf{f}_t$ until $t = 10,000$. It can be observed that the increase of execution time is almost linear w.r.t. the increasing of d for all the compared approaches. Moreover, it is obvious that USO is more efficient than the counterparts in Fig. 4. This is because USO forms and searches fewer non-overlapping neighborhood sets, and obtains a more concise feature subset. For Fig. 5, we fix d at 1000 with $d_n = d_c = 500$ and still let $\mathsf{v}^t \in \{3, 4, 5, 6\}$ for categorical features, to randomly generate data objects until $n = 10,000$. It can be seen that with the linear increase of n, the execution time for selecting 1000 features increases polynomially. All the above observations conform with the time complexity analysis at the end of Section III-C.

V. Concluding Remarks

This paper has proposed a novel OHFS approach called USO for boosting the analysis of real complex biomedical data sets. USO mainly addresses the two difficulties in OHFS, i.e., heterogeneity caused by the mixed features and dynamic

feature space due to streaming features, by proposing (1) SAN-R, a flexible neighborhood relation for forming neighborhood sets, and (2) UDM, a unified distance metric as the basis of (1). To the best of our knowledge, this is the first attempt to simultaneously tackle the online and heterogeneity issues in feature selection, which is more challenging than only coping with one of them. The superiority of USO has been well demonstrated by the comparative results, significance tests, ablation study, etc., on various real public biomedical data sets. Moreover, USO is parameter-free, efficient, and interpretable.

Due to space limitation, we have to omit case studies and evaluation of feature input order robustness of USO in this paper. It is worth noting that the omitted results have indeed confirmed that USO is very robust to different feature orders. These results and discussions will be available soon in a journal version of this paper. Moreover, considering the challenging issues under the OHFS settings, e.g., streaming data with concept drifts [36], time-series data with missing values [37], and very large-scale data [38], would be promising future research directions.

REFERENCES

[1] L. Sun, Z. Mo, and et al., "Adaptive feature selection guided deep forest for covid-19 classification with chest ct," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2798–2805, 2020.

[2] E. S. Walsh, H. T. Ghashghaei, and X. Peng, "Feature selection using co-occurrence correlation improves cell clustering and embedding in single cell rnaseq data," in *Proceedings of the 2021 International Conference on Bioinformatics and Biomedicine*, 2021, pp. 751–756.

[3] U. Pale, T. Teijeiro, and et al., "Exg signal feature selection using hyperdimensional computing encoding," in *Proceedings of the 2022 International Conference on Bioinformatics and Biomedicine*, 2022, pp. 1688–1693.

[4] Y.-m. Cheung and H. Zeng, "Local kernel regression score for selecting features of high-dimensional data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 12, pp. 1798–1802, 2009.

[5] N. AlNuaimi, M. M. Masud, M. A. Serhani, and et al., "Streaming feature selection algorithms for big data: A survey," *Applied Computing and Informatics*, vol. 18, no. 1/2, pp. 113–135, 2020.

[6] A. Zeng, H. Rong, D. Pan, and et al., "Discovery of genetic biomarkers for alzheimers disease using adaptive convolutional neural networks ensemble and genome-wide association studies," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 13, no. 4, pp. 787–800, 2021.

[7] X. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu, "Online feature selection with streaming features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1178–1192, 2012.

[8] Y. Zhang, "Advances in categorical data clustering," Ph.D. dissertation, Hong Kong Baptist University, 2019.

[9] Y. Zhang and Y.-m. Cheung, "Exploiting order information embedded in ordered categories for ordinal data clustering," in *Proceedings of the 24th International Symposium on Methodologies for Intelligent Systems*, 2018, pp. 247–257.

[10] Y. Zhang, Y.-m. Cheung, and K. C. Tan, "A unified entropy-based distance metric for ordinal-and-nominal-attribute data clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 1, pp. 39–52, 2019.

[11] Y. Zhang and Y.-m. Cheung, "An ordinal data clustering algorithm with automated distance learning," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

[12] L. Zhao, Y. Zhang, Y. Ji, A. Zeng, F. Gu, and X. Luo, "Heterogeneous drift learning: Classification of mix-attribute data with concept drifts," in *Proceedings of the 9th International Conference on Data Science and Advanced Analytics*, 2022, pp. 1–10.

[13] Y. Zhang and Y.-m. Cheung, "A new distance metric exploiting heterogeneous interattribute relationship for ordinal-and-nominal-attribute data clustering," *IEEE Transactions on Cybernetics*, vol. 52, no. 2, pp. 758–771, 2020.

[14] S. Eskandari and M. M. Javidi, "Online streaming feature selection using rough sets," *International Journal of Approximate Reasoning*, vol. 69, pp. 35–57, 2016.

[15] M. M. Javidi and S. Eskandari, "Online streaming feature selection: a minimum redundancy, maximum significance approach," *Pattern Analysis and Applications*, vol. 22, pp. 949–963, 2019.

[16] C.-O. J. A. Solorio-Fernández, S. and et al., "A survey on feature selection methods for mixed data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 55, no. 4, pp. 2821–2846, 2022.

[17] Q. Hu, J. Liu, and D. Yu, "Mixed feature selection based on granulation and approximation," *Knowledge-Based Systems*, vol. 21, no. 4, pp. 294–304, 2008.

[18] P. Zhang, T. Li, Z. Yuan, and et al., "Heterogeneous feature selection based on neighborhood combination entropy," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[19] Y. Zhang and Y.-m. Cheung, "Discretizing numerical attributes in decision tree for big data analysis," in *Proceedings of the 2014 International Conference on Data Mining Workshop*, 2014, pp. 1150–1157.

[20] S. Sharmin, M. Shoyaib, A. Ali, M. Khan, and O. Chae, "Simultaneous feature selection and discretization based on mutual information," *Pattern Recognition*, vol. 91, pp. 162–174, 2019.

[21] Y. Zhang and Y.-m. Cheung, "Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3560–3576, 2022.

[22] Y. Zhang, Y.-m. Cheung, and A. Zeng, "Het2hom: representation of heterogeneous attributes into homogeneous concept spaces for categorical-and-numerical-attribute data clustering," in *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, 2022, pp. 3758–3765.

[23] Y. Zhang and Y.-M. Cheung, "Graph-based dissimilarity measurement for cluster analysis of any-type-attributed data," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[24] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, p. 99, 2000.

[25] P. Zhou, P. Li, S. Zhao, and et al., "Online early terminated streaming feature selection based on rough set theory," *Applied Soft Computing*, vol. 113, p. 107993, 2021.

[26] P. Zhou, X. Hu, P. Li, and et al., "Online streaming feature selection using adapted neighborhood rough set," *Information Sciences*, vol. 481, pp. 258–279, 2019.

[27] ——, "Ofs-density: A novel online streaming feature selection method," *Pattern Recognition*, vol. 86, pp. 48–61, 2019.

[28] Y. Lv, Y. Lin, X. Chen, and et al., "Online streaming feature selection based on feature interaction," in *Proceedings of the 11th International Conference on Knowledge Graph*, 2020, pp. 49–57.

[29] S. Li, K. Zhang, Y. Li, and et al., "Online streaming feature selection based on neighborhood rough set," *Applied Soft Computing*, vol. 113, p. 108025, 2021.

[30] D. Dua and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[31] "CLOPINET: Feature selection challenge, NIPS 2003," http://clopinet.com/isabelle/Projects/NIPS2003/.

[32] "CLOPINET: Performance prediction challenge, WCCI 2006," http://clopinet.com/isabelle/Projects/modelselect/.

[33] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[34] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[35] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[36] M. Zhao, Y. Zhang, Y. Ji, and Y. Lu, "Unsupervised concept drift detection via imbalanced cluster discriminator learning," in *Proceedings of the 6th Chinese Conference on Pattern Recognition and Computer Vision*, 2023, pp. 1–12.

[37] Z. Zhang, Y. Zhang, A. Zeng, D. Pan, Y. Ji, Z. Zhang, and J. Lin, "Time-series data imputation via realistic masking-guided tri-attention bi-gru," in *Proceedings of the 26th European Conference on Artificial Intelligence*, 2023, pp. 3074–3082.

[38] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: a review," *Complex & Intelligent Systems*, vol. 8, no. 3, pp. 2663–2693, 2022.