


Feature Selection Using Neighborhood based Entropy


Fatemeh Farnaghi-Zadeh

(Department of Computer Engineering, Faculty of Engineering, Arak University, Arak
38156-8-8349, Iran

 <https://orcid.org/0000-0002-4940-1302>, s39713161005@msc.araku.ac.ir)


Mohsen Rahmani

(Department of Computer Engineering, Faculty of Engineering, Arak University, Arak
38156-8-8349, Iran

 <https://orcid.org/0000-0001-6890-192X>, m-rahmani@araku.ac.ir)

Maryam Amiri

(Department of Computer Engineering, Faculty of Engineering, Arak University, Arak
38156-8-8349, Iran

 <https://orcid.org/0000-0002-7411-9552>, m-amiri@araku.ac.ir)

Abstract: Feature selection plays an important role as a preprocessing step for pattern recognition and machine learning. The goal of feature selection is to determine an optimal subset of relevant features out of a large number of features. The neighborhood discrimination index (NDI) is one of the newest and the most efficient measures to determine distinguishing ability of a feature subset. NDI is computed based on a neighborhood radius (ϵ). Due to the significant impact of ϵ on NDI, selecting an appropriate value of ϵ for each data set might be challenging and very time-consuming. This paper proposes a new approach based on target PointS To compute neighborhood relations (EPSTEIN). At first, all the data points are sorted in the descending order of their density. Then, the highest density data points are selected as many as the number of classes. To determine the neighborhood relations, the circles centered on the target points are drawn and the points inside or on the circles are considered to be neighbors. In the next step, the significance of each feature is computed and a greedy algorithm selects appropriate features. The performance of the proposed approach is compared to both the commonest and newest methods of feature selection. The experimental results show that EPSTEIN could select more efficient subsets of features and improve the prediction accuracy of classifiers in comparison to the other state-of-the-art methods such as Correlation-based Feature Selection (CFS), Fast Correlation-Based Filter (FCBF), Heuristic Algorithm Based on Neighborhood Discrimination Index (HANDI), Ranking Based Feature Inclusion for Optimal Feature Subset (KNFI), Ranking Based Feature Elimination (KNFE) and Principal Component Analysis and Information Gain (PCA-IG).

Keywords: Feature Selection, Discrimination Index, Neighborhood Relations, Density, Entropy, Distinguishing Ability

Categories: H.1.1, H.2.8, I.5.2

DOI: 10.3897/jucs.79905

1 Introduction

Data sets today are described by a large number of features, which might contain redundant or irrelevant features [Liu et al., 2017, Armanfard et al., 2015, Wang et al.,

2017, Amiri et al., 2020]. These unnecessary features might lead to the curse of dimensionality [Liu et al., 2017], increase in the classification error and training time of algorithms and the problem of over-fitting [Liu et al., 2019]. Feature selection is a preprocessing step to reduce the dimension of data [Fard et al., 2013]. Feature selection is a technique to determine an optimal subset of features with strong classification ability according to certain assessment criteria in a way that data analysis is simplified and high-dimensional characteristics are acquired by analyzing low-dimensional data [Wang et al., 2017].

In general, feature selection methods proposed until now could be divided into three groups [Liu et al., 2017, Liu et al., 2019, Chen and Chen, 2015, Dong and Liu, 2018]: the filter model, the wrapper model and the hybrid (embedded) model: 1) The filter model ranks independently features using some score measures and selects the top-ranked ones. The score is based on information theory and statistics and basically measures the correlation between the feature and the decision attribute. Indeed, the filter approaches work independently of the predictor. Although the computation cost and generalization ability of these methods are low and high respectively, their main limitation is to ignore feature redundancy and feature interdependencies [Kamalov and Thabtah, 2017, Yu and Liu, 2004]. 2) In the wrapper model, the whole powerset of the feature set is considered and for each candidate subset, the generalization error is computed. In fact, in the wrapper model, the performance of feature selection depends on the classifier directly [Gaudel and Sebag, 2010, Mafarja and Mirjalili, 2018, Yang and Ong, 2011]. 3) The hybrid model performs feature selection in the training phase [Gaudel and Sebag, 2010]. So it needs a specific learning algorithm before conducting feature selection in the process of training [Liu et al., 2017, Fard et al., 2013].

Entropy and neighborhood relations have an important role in pattern recognition and data mining. Hence many methods have been proposed based on them [Wang et al., 2017, Dai et al., 2012, Battiti, 1994, Hu et al., 2011, Hu et al., 2008b, Zhu and Hu, 2013]. Entropy, one of the most prominent approaches in information theory [Cicioğlu, 2021], is an uncertainty measure that characterizes the distinguishing information of an arbitrary subset of features [Shannon, 2001]. As the conditional entropy of the decision attribute on a subset of features decreases, the subset has more ability to distinguish samples with different class labels. Given the samples characterized by numerical features, the neighborhood can be used to distinguish them and extract similarity classes from them. The neighborhood relations induced by a feature subset have an important impact on its distinguishing information [Wang et al., 2017]. In [Wang et al., 2017], the neighborhood relation is computed based on a neighborhood radius, called ϵ . Since the value of ϵ has a significant effect on the neighborhood relation, selecting an appropriate value for it is very challenging.

This paper proposes a new approach based on target Points To compute neighborhood relations (EPSTEIN). In EPSTEIN, the data points are ranked based on density-based criteria to establish the neighborhood relationship. Then, according to the number of classes in the data set, high-ranking data points are selected as target points, and neighborhood relations are defined around these points. The proposed approach is based on a parameter, called σ , which determines what percentage of the data points is considered to be neighbors to compute the local density. Finally, based on a significance measure, a greedy algorithm selects a subset of features. So, the contributions of this paper are as follows:

- This paper presents a new approach based on target points for neighborhood relations construction (Section 3.1).
- According to our experimental results, the parameter of EPSTEIN has a small

impact on the neighborhood relation. In other words, EPSTEIN does not need to compute the radius neighborhood, which has a significant impact on the quality of the neighborhood relations (Section 4.1).

- As our experimental results show, EPSTEIN could select an appropriate subset of features and improve the accuracy of classification algorithms (Section 4.2).

The rest of the paper is organized as follows: Section 2.1 reviews related works on feature selection. In section 3, the proposed approach is described. We present the experimental results in Section 4. Finally, the paper is concluded with our future work in Section 5.

2 Related Works

In general, feature selection methods could be divided into three groups [Liu et al., 2017]: the wrapper model, the embedded model, and the filter model. In section 2.1, we investigate these methods. In addition, since EPSTEIN selects features using neighborhood based entropy, we review the works that focus on the entropy defined on the neighborhood in section 2.2.

2.1 Feature Selection Approaches

Figure 1 shows the structure of the three models. As the figure shows, the filter model ranks independently features using some score measures and selects the top-ranked ones. The wrapper model explores the feature space and considers the candidate subsets of the feature set and evaluates the performance of each subset based on the classification error. The hybrid model is also a combination of the filter model and the wrapper model. In the following subsections, the newest and the most prominent works of each group are reviewed briefly.

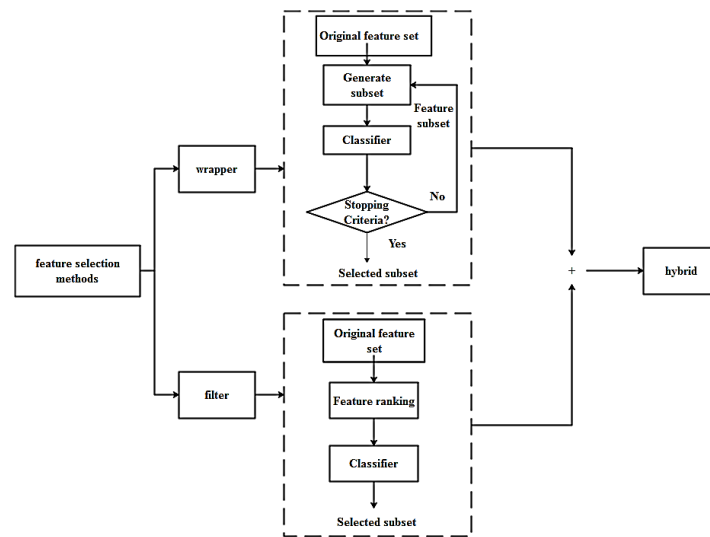


Figure 1: The structure of the different models of feature selection

2.1.1 Filter models

Hall in [Hall, 2000] proposes Correlation-based Feature Selection (CFS). At first, CFS computes a matrix of feature-class and feature-feature correlations from the training data. In the next step, it searches the feature subset space using a best-first search algorithm. A subset of features, that are highly correlated with the class attribute and have little correlation with each other, is selected.

Relief is proposed in [Kira and Rendell, 1992]. It randomly selects samples from a data set and updates the significance of each feature based on the difference between the selected instance and the two closest objects of the same class (near-hit) and the opposite class. Relief considers the difference in the values of a feature for the two nearest neighbors of the same class and the difference between the values of the feature for the objects of different classes. Based on the difference values, the importance of features adjusts. In [Kononenko, 1994], an improved version of Relief, called ReliefF, is extended to handle multi-class data sets and noise. It updates the weights based on the k-near hits and the k-near misses for each of the classes that are different from the class of the random sample. The challenge of this method is to determine the threshold value for selecting weighted features.

In [Yu and Liu, 2003], an approximation method for relevance and redundancy analysis, called FCBF (Fast Correlation-Based Filter), is presented. Firstly it selects a subset of relevant features, and then selects predominant features from relevant ones. FCBF ranks features according to their relevance to the class. Although the method is faster and more efficient than ReliefF and CFS, in some cases it has errors in recognizing redundant features.

The minimal-redundancy-maximum-relevance (mRMR) method [Peng et al., 2005] employs mutual information (MI) to evaluate feature relevance and feature redundancy. The mRMR function uses the SFS search strategy [Pudil et al., 1994] to build the best feature subset. The subset of selected features is initially empty. At each step, the best feature based on the evaluation criteria is added to this subset. The principal disadvantages of this algorithm are to determine the number of the features that should be selected and the size of the optimal solution.

Thabtah et al. in [Pudil et al., 1994] propose a feature selection method called Least Loss (L2) that significantly reduces the dimensionality of data. It disposes weakly correlated variables without diminishing the predictive performance of classifiers. The greater the value of L2 is, the more relevant a feature is to the target class. The main advantage of the L2 measure is its simplicity, ease of understanding, and intuitiveness. The challenge of this method is to determine the cut-off value.

A feature selection method based on a correlation measure between continuous and discrete features (ECMBF) is proposed in [Jiang and Wang, 2016]. The method removes weakly relevant and irrelevant features, as well as relevant but redundant features. The performance of this approach might be influenced by the number of training samples.

Mariello et al. in [Mariello and Battiti, 2018] propose a new algorithm named NEFS for filtering features that maximize MI between the selected subset and the class variable and, at the same time, tries to minimize MI between the selected features. To overcome the limitation of the traditional pairwise MI estimators, a new MI-related measure is introduced, which can be applied to multiple features. The time complexity of the algorithm is high.

In [Wang et al., 2017], a Heuristic Algorithm Based on Neighborhood Discrimination Index (HANDI) is proposed for feature selection. The neighborhood discrimination index (NDI) is proposed to characterize the distinguishing information of a neighborhood

relation. It reflects the distinguishing ability of a feature subset. As the conditional discrimination index of a feature subset gets smaller, the distinguishing ability of the feature subset gets greater and hence, the feature subset would be more important. Based on the discrimination measure, the significance measure of a candidate feature is defined and a greedy forward algorithm is suggested for feature selection. Since the neighborhood radius (ϵ) has a great effect on calculating the distinguishing ability of feature subsets and depends on data sets, determining the value of ϵ is challenging. Omuya et al. in [Omuya et al., 2021] develop a hybrid filter model for feature selection based on Principal Component Analysis and Information Gain (PCA-IG). The model applies PCA that employs feature correlation at the initial level and IG that uses entropy evaluation at the second level.

2.1.2 Wrapper models

In [Aghdam et al., 2009], to improve the performance of text categorization, a feature selection algorithm based on ant colony optimization [Bonabeau et al., 1999] is proposed. The algorithm is easily implemented and since it uses a simple classifier, its computational complexity is very low. The disadvantage of this algorithm is uncertain convergence time.

Bouaguel in [Bouaguel, 2016] proposes a new wrapper feature selection method for big data. It is based on the random search and the genetic algorithm. In the first step, fewer redundant features are selected without sacrificing quality. In the next step, subsets are generated by using the genetic algorithm (GA) [Goldberg, 1989]. The method requires prior knowledge about data sets.

Sahebi et al. in [Sahebi et al., 2020] propose a generalized wrapper-based feature selection, called GeFeS, which is based on a parallel intelligent genetic algorithm. To validate the learning model, the authors propose a new operator for weighting features, improve the mutation and crossover operators, and integrate nested cross-validation into the GA process. Although GeFeS works on different data sets, it requires prior knowledge about them.

2.1.3 Hybrid models

In [Suresh and Narayanan, 2019], a hybrid feature selection approach is presented that incorporates the benefits of both filter and wrapper methods. At first, the filter part ranks features. In the next step, for subset selection, the first ranked feature is added to the empty subset and classification accuracy is evaluated. Then, the next ranked features are considered in order. A new feature will be inserted into the feature subset only when it improves the classification accuracy compared to the previous result. Different methods of feature ranking and subset selection are used in the algorithm. Selecting the best filter and wrapper method is challenging.

Wei et al. in [Wei et al., 2020] suggest a feature selection algorithm, called Dynamic Feature Importance-based Feature Selection (DFIFS). It dynamically selects features according to their Dynamic Feature Importance (DFI) index in the selection process. DFI is defined based on both feature redundancy and feature importance. By combining DFIFS and mRMR, the authors propose a hybrid method called M-DFIFS. mRMR is used to filter out redundant or irrelevant features, while DFIFS is applied to adjust the selected feature subset. The performance of this algorithm depends on the classifier.

Thejas et al. in [Thejas et al., 2019] propose a feature selection mechanism based on

the combination of the filter and the wrapper techniques. They cluster the data using mini-batch K-means clustering and rank them using normalized mutual information. Then, a greedy search method by using Random Forest is applied to get the optimal set of features. They propose two approaches for the selection of features:

- MiniBatch K-means Normalized Mutual Information Feature Inclusion (KNFI): the ranked features from the first phase are added one by one into the subset. If the addition of the features enhances the classification accuracy, the feature is added or else it is discarded. This process loops for all the features.
- Mini-Batch K-means Normalized Mutual Information least ranked Feature Exclusion (KNFE): this is a linear elimination approach where the least ranked features are eliminated one by one from the entire set of the features. Initially, the list consists of all the features and the classification accuracy is calculated for the entire list. Then, in every loop, one least ranked feature is removed from the list. This process is repeated until the list becomes empty. The highest performance among all the iterations is considered as the outcome.

2.2 Neighborhood Entropy

Neighborhood entropy proposed in [Mariello and Battiti, 2018] overcomes the limitation of traditional pairwise MI estimators and can be applied to multiple features. This score does not involve the explicit estimation of probability distributions or the computation of local and global affinity matrices of the graph-based methods, and can be used with features of integer or real values. The conditional class entropy evaluated on the neighborhoods of the points can be used as a relevance score for selecting the most informative features. Even in the situation of nonuniform distributions, the previous observations remain valid if one considers neighborhoods with a fixed number of neighbors instead of a fixed radius.

The concept of neighborhood entropy is defined to measure the uncertainty of numerical data. When the classification performance of the original data set is poor, the corresponding evaluation functions have lower measured values; thus, monotonic attribute reduction methods cannot obtain great reduction results [Li et al., 2013]. To address this issue, in [Sun et al., 2019a] some concepts of neighborhood entropy-based uncertainty measures are proposed to investigate the uncertainty of knowledge in neighborhood decision systems. Since the Fisher score method occasionally selects redundant attributes, which affects the classification result [Hasanloei et al., 2018], the Fisher score with neighborhood rough sets is combined to reduce the initial dimensions and improve the classification performance of high-dimensional gene expression data sets.

The gene selection method proposed in [Sun et al., 2019b] is based on the filter approach, in which a heuristic search algorithm is used to find an optimal gene subset with neighborhood rough sets for the gene expression data. Since the information entropy is not suitable for measuring the neighborhood class in the numeric data sets, the concept of neighborhood is combined with information theory measures. The neighborhood rough sets and entropy measure-based gene selection with Fisher score for tumor classification are proposed. Firstly, the Fisher score method is employed to eliminate irrelevant genes to significantly reduce computation complexity. Next, some neighborhood entropy-based uncertainty measures are investigated for handling the uncertainty and noisy of gene expression data. Finally, a joint neighborhood entropy-based gene selection algorithm with the Fisher score is presented to improve the classification performance of gene expression data.

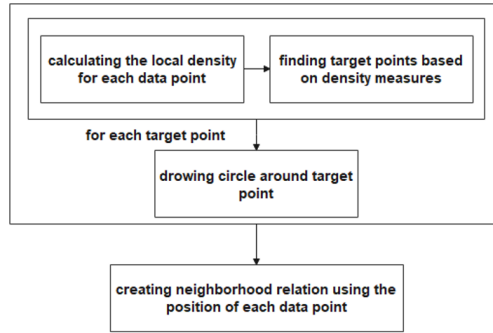


Figure 2: The steps of neighborhood relations construction in EPSTEIN

3 Foundation of EPSTEIN

In EPSTEIN, the neighborhood relation is defined around the target points. A target point is a point whose local density is high, the distance from points with higher densities is large and the distance from ones with lower densities is small [Du et al., 2016, Li and Tang 2018, Pourbahrami et al., 2019]. All the data points are sorted in the descending order of their density. Then, the target points are selected as many as the number of classes. Figure 2 shows the steps of neighborhood relations construction. Finally, based on a significance measure, a greedy algorithm selects a subset of features. The selected subset of features is empty firstly. In each iteration, the most significant feature is added to the subset. The algorithm terminates when the significance of any remaining feature is less than a threshold.

In the following subsections, EPSTEIN is described in detail: In section 3.1, the background concepts such as **target point**, **local density** and **score** are defined and a new neighborhood relation is introduced. Then, section 3.2 investigates the measure which is employed to rank features. Finally, in section 3.3, the algorithm used to select features is explained. Table 1 shows the sample data set used to introduce concepts in the following examples. The sample data set contains 7 instances, each described by two features a and b . The instances are also classified into two classes *Class 1* and *Class 2*.

Table 1: The sample data set

No	Feature		Class
	a	b	
1	8.2	16.3	1
2	7.7	15.56	1
3	8.1	114.8	1
4	8.82	15.1	1
5	10.9	18	2
6	11.3	18.8	2
7	12	19.1	2

Table 2: The description of the parameters/variables of Algorithm 1

Variable/Parameter	Description
S	Data Set
A	Feature Set
D	Decision Attribute
B	Unselected Features
<i>SelectedFeatures</i>	Selected Features
σ	The percentage of the data points to compute density
α	The threshold to identify the significant features

3.1 Neighborhood relations

The concept of neighborhood plays an important role in numerical spaces. The subset of the samples which have the similar feature values can be identified by using the

neighborhood relations computed based on distance [Hu et al., 2011, Hu et al., 2008a]. In this paper, a new neighborhood relation is constructed based on local density and the circles centered on the target points. For this purpose, the target points should be determined firstly. The target points are identified by density based measures. The density of a data point implies how many data points congregate around it. In a similar way to [Du et al., 2016, Li and Tang 2018, Pourbahrami et al., 2019], the local density of each data point is defined as follows. Note that the sign $|\cdot|$ is used to denote the cardinality of a set or relation [Wang et al., 2017].

Definition 1. Given the data set S and the data point $x \in S$, the number of the nearest neighbors of x is $r = \lceil \sigma \times |S| \rceil$ where σ is the percentage of the data points that should be considered to be neighbors and $\lceil \cdot \rceil$ is the ceiling function.

Definition 2. Given the data set S and the data point $M_i = \{M_{1i}, M_{2i}, \dots, M_{mi}\} \in S$ where m is the number of the features, the local density of the data point M_i , ρ_i , is defined as Eq. 3.1:

$$\rho_i = \exp\left(-\frac{1}{r} \sum_{M_j \in N(M_i)} d(M_i, M_j)^2\right) \quad (3.1)$$

Where r is the number of the neighbors of M_i , $N(M_i)$ is r nearest neighbors of the data point M_i and $d(M_i, M_j)$ is the Euclidean distance between the data points M_i and M_j .

ρ has an important role in identifying the target points. When the local density of a point is high, it means that more points concentrate around it. Then, it is more probable that a neighborhood circle is centered on that point.

Example 1. Consider the sample data set in Table 1. Assume $\sigma = 0.01$ (or 1%). To compute the local density of each data point, firstly, the matrix *DistMatrix*, which contains the Euclidean distance between each pair of the data points, is computed as follows:

$$DistMatrix = \begin{bmatrix} 0 & 0.207694 & 0.349612 & 0.314117 & 0.742003 & 0.926154 & 1.097714 \\ 0.207694 & 0 & 0.199729 & 0.281578 & 0.935844 & 1.126350 & 1.295280 \\ 0.349612 & 0.199729 & 0 & 0.181395 & 0.988851 & 1.191279 & 1.350040 \\ 0.314117 & 0.281578 & 0.181395 & 0 & 0.829956 & 1.035874 & 1.188379 \\ 0.742003 & 0.935844 & 0.988851 & 0.829956 & 0 & 0.208006 & 0.361776 \\ 0.926154 & 1.126350 & 1.191279 & 1.035874 & 0.208006 & 0 & 0.177111 \\ 1.097714 & 1.295280 & 1.350040 & 1.188379 & 0.361776 & 0.177111 & 0 \end{bmatrix}$$

Since $\sigma = 0.01$, we have $r = \lceil 0.01 \times 7 \rceil = 1$. Then, based on Eq. (3.1), the density of each data point is computed:

$$\begin{aligned} \rho_1 &= e^{-(0.207694)^2} = 0.95778033 & \rho_2 &= e^{-(0.199729)^2} = 0.96089337 \\ \rho_3 &= e^{-(0.181695)^2} = 0.96763118 & \rho_4 &= e^{-(0.181395)^2} = 0.96763118 \\ \rho_5 &= e^{-(0.208006)^2} = 0.95765602 & \rho_6 &= e^{-(0.177111)^2} = 0.96911857 \\ \rho_7 &= e^{-(0.177111)^2} = 0.96911857 \end{aligned}$$

Definition 3. Given the data point M_i , δ_i is the distance from M_i to the nearest neighbor whose local density is greater than ρ_i [Du et al., 2016, Li and Tang 2018, Pourbahrami et al., 2019]:

$$\delta_i = \begin{cases} \min_{j: \rho_i < \rho_j} \{d(M_i, M_j)\}, & \text{if } \exists M_j \in S, \rho_i < \rho_j \\ \max_j \{d(M_i, M_j)\}, & \text{otherwise} \end{cases} \quad (3.2)$$

As Eq. 3.2 shows, if the local density of M_i (ρ_i) is greater than or equal to all the data points', δ_i is the maximum distance between M_i and the other data points.

Definition 4. Given the data point M_i , τ_i is the distance from M_i to the nearest neighbor whose local density is less than ρ_i [Li and Tang 2018]:

$$\tau_i = \begin{cases} \delta_i, & \text{if } \forall M_j \in S, \rho_i \leq \rho_j \\ \min_{j: \rho_i > \rho_j} \{d(M_i, M_j)\}, & \text{otherwise} \end{cases} \quad (3.3)$$

According to Eq. 3.3, if ρ_i is less than or equal to all the data points', τ_i is set to δ_i .

Definition 5. Given the data point M_i , the score of M_i is computed as follows [Li and Tang 2018]:

$$\text{score}(M_i) = \rho_i \times (\delta_i - \tau_i) \quad (3.4)$$

To identify the target points, the data points are ranked based on the score. The data points with the highest score are the target points. Therefore, as Eq. 3.4 shows, the target points have three key attributes: 1) they have large density (ρ_i), 2) their distance from points whose density is greater than themselves (δ_i) is large and 3) their distance from points whose density is less than themselves (τ_i) is small.

Definition 6. Let C be the number of classes. The C data points with the highest scores are **target points**.

Example 2. Consider Example 1 again. Firstly, for each data point, δ and τ are computed.

$$\begin{aligned} \delta_1 &= \min(d(1, 2), d(1, 3), d(1, 4), d(1, 6), d(1, 7)) = 0.2077 \\ \delta_2 &= \min(d(2, 3), d(2, 4), d(2, 6), d(2, 7)) = 0.1997 \\ \delta_3 &= \min(d(3, 6), d(3, 7)) = 1.1913 \\ \delta_4 &= \min(d(4, 6), d(4, 7)) = 1.0359 \\ \delta_5 &= \min(d(5, 1), d(5, 2), d(5, 3), d(5, 4), d(5, 6), d(5, 7)) = 0.2080 \\ \delta_6 &= \max(d(6, 1), d(6, 2), d(6, 3), d(6, 4), d(6, 5), d(6, 7)) = 1.1913 \\ \delta_7 &= \max(d(7, 1), d(7, 2), d(7, 3), d(7, 4), d(7, 5), d(7, 6)) = 1.35 \\ \tau_1 &= \min(d(1, 5)) = 0.742 \quad \tau_2 = \min(d(2, 1), d(2, 5)) = 0.2077 \\ \tau_3 &= \min(d(3, 1), d(3, 2), d(3, 5)) = 0.1997 \quad \tau_4 = \min(d(4, 1), d(4, 2), d(4, 5)) = 0.2816 \\ \tau_5 &= \delta_5 = 0.2080 \quad \tau_6 = \min(d(6, 1), d(6, 2), d(6, 3), d(6, 4), d(6, 5)) = 0.2080 \\ \tau_7 &= \min(d(7, 1), d(7, 2), d(7, 3), d(7, 4), d(7, 5)) = 0.3618 \end{aligned}$$

Based on Eq. 3.4, *score* is computed as follows:

$$\begin{aligned} \text{score}_1 &= \rho_1 \times (\delta_1 - \tau_1) = -0.5117 & \text{score}_2 &= \rho_2 \times (\delta_2 - \tau_2) = -0.0076 \\ \text{score}_3 &= \rho_3 \times (\delta_3 - \tau_3) = 0.9594 & \text{score}_4 &= \rho_4 \times (\delta_4 - \tau_4) = 0.7299 \\ \text{score}_5 &= \rho_5 \times (\delta_5 - \tau_5) = 0 & \text{score}_6 &= \rho_6 \times (\delta_6 - \tau_6) = 0.9529 \\ \text{score}_7 &= \rho_7 \times (\delta_7 - \tau_7) = 0.9577 \end{aligned}$$

Since $C = 2$, the two data points M_3 and M_7 are the target points.

After the target points are determined, the circles centered on them are drawn for forming the neighborhood relations.

Definition 7. Given the data set S , the set of the target points $T = \{T_1, T_2, \dots, T_C\}$ where C is the number of classes and $M = S - T$, the radius of the circles centered at each T_t , $1 \leq t \leq C$, is:

$$R_{T_t} = \max\{d(T_t, M_i) | M_i \in M, d(T_t, M_i) < \min_{l=1}^C d(T_l, M_i), t \neq l\} \quad (3.5)$$

Definition 8. Given the data set S , the set of the target points $T = \{T_1, T_2, \dots, T_C\}$ where C is the number of classes and $M = S - T$, the furthest points from each T_t , $1 \leq t \leq C$ is:

$$FP_{T_t} = \{M_i | M_i \in M, d(T_t, M_i) = R_{T_t}\} \quad (3.6)$$

According to Eqs. 3.5 and 3.6, the furthest points from the target point T_t are the points whose distance from the other target points is greater than T_t 's. The points whose distance from the target point is greater than R_{T_t} , are not in the neighborhood of T_t . After drawing the circles centered at the target points, the neighborhood relation is determined. The neighborhood relation is represented by a neighborhood matrix.

Definition 9. Given $|S| = n$ and $1 \leq i, j \leq n$, the density based neighborhood matrix $R^p = (r_{ij})_{n \times n}$ is defined as follows:

$$r_{ij} = \begin{cases} 1, & \text{if } i = j \text{ or } M_i \text{ and } M_j \text{ are in the same circle or on the same circle} \\ 0, & \text{otherwise} \end{cases} \quad (3.7)$$

Based on Eq. 3.7, the points inside or on the circles are considered to be neighbors of each other. The points falling inside the overlap area of the circles or out of their radius are considered to be the neighbor of their nearest target points.

Example 3. Consider Example 2 again. Since the points M_3 and M_7 are the target point, we have $T = \{3, 7\}$. Based on Eqs. 3.5 and 3.6, we have:

$$\begin{aligned} R_{T_3} &= \max(d(3, 1), d(3, 2), d(3, 4)) = d(3, 1) = 0.3496 & FP_{T_3} &= 1 \\ R_{T_7} &= \max(d(7, 6), d(7, 5)) = d(7, 5) = 0.3618 & FP_{T_7} &= 5 \end{aligned}$$

In the next step, as Fig. 3 shows, the circles centered at the target points are drawn and R^p is constructed.

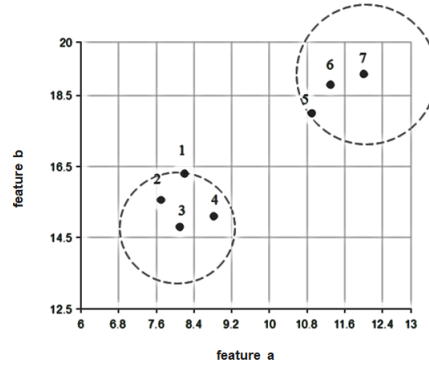


Figure 3: The neighborhood matrix R^p and the circles centered at the target points

3.2 Feature importance

In [Wang et al., 2017], the conditional discrimination index and NDI are computed by using the neighborhood relations. The neighborhood relations are based on the neighborhood radius ϵ . In this paper, inspired by [Wang et al., 2017], the two indexes are defined based on EPSTEIN.

Definition 10. Let B_1 and B_2 be two subsets of features, $|S| = n$, $R_{B_1}^\rho$, $R_{B_2}^\rho$ be two neighborhood relations induced by B_1 , B_2 based on EPSTEIN, respectively. The neighborhood discrimination index B_1 and the conditional discrimination index of B_1 on B_2 are defined as Eqs. 3.8 and 3.9 respectively:

$$H^\rho(B_1) = \log \frac{n^2}{|R_{B_1}^\rho|} \quad (3.8)$$

$$H^\rho(B_1|B_2) = \log \frac{|R_{B_2}^\rho|}{|R_{B_1}^\rho \cap R_{B_2}^\rho|} \quad (3.9)$$

In Eq. 3.8, as $|R_{B_1}^\rho|$ increases, NDI decreases, which implies that the distinguishing ability of B_1 is low. In Eq. 3.9, if a is a feature and $B_2 = B_1 \cup \{a\}$, then $H^\rho(B_1|B_2)$ shows how much the feature a can change the distinguishing ability of B_1 . If B is a subset of features and D is the decision attribute, $H(D|B)$ indicates the ability of B to distinguish samples with different class labels. The smaller the $H(D|B)$, the greater the distinguishing ability of B . Adding a feature to B can increase or decrease the distinguishing ability of samples with different class labels. If a decreases the distinguishing ability, it is called redundant. In a similar way to [Pudil et al., 1994], the significant degree of features is defined as follows.

Definition 11. Let A be the feature set, $B \subseteq A$, $a \in A - B$ and D be the decision attribute. The significant degree of the feature a with respect to B and D is:

$$SIG(a, B, D) = H^\rho(D|B) - H^\rho(D|B \cup \{a\}) \quad (3.10)$$

Note that if $B = \emptyset$, then $H^\rho(D|B) = H^\rho(D)$.

According to Eq. 3.10, a larger value of $SIG(a, B, D)$ implies that a is more prominent for D .

Example 4. Assume A is the feature set, $B \subseteq A$, $a, b \in A - B$, D is the decision attribute and R_B^ρ , $R_{B \cup \{a\}}^\rho$, $R_{B \cup \{b\}}^\rho$ and R_D (the decision equivalence relation) are as follows:

$$R_B^\rho = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad R_{B \cup \{a\}}^\rho = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad R_{B \cup \{b\}}^\rho = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad R_D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$SIG(a, B, D)$ and $SIG(b, B, D)$ are computed as follows:

$$\begin{aligned} H^\rho(D|B) &= \log \frac{|R_B^\rho|}{|R_D \cap R_B^\rho|} = \log \frac{5}{3} = 0.7369 \\ H^\rho(D|B \cup \{a\}) &= \log \frac{|R_{B \cup \{a\}}^\rho|}{|R_D \cap R_{B \cup \{a\}}^\rho|} = \log \frac{5}{3} = 0.7369 \\ H^\rho(D|B \cup \{b\}) &= \log \frac{|R_{B \cup \{b\}}^\rho|}{|R_D \cap R_{B \cup \{b\}}^\rho|} = \log \frac{7}{3} = 1.2223 \\ SIG(a, B, D) &= H^\rho(D|B) - H^\rho(D|B \cup \{a\}) = 0.7369 - 0.7369 = 0 \\ SIG(b, B, D) &= H^\rho(D|B) - H^\rho(D|B \cup \{b\}) = 0.7369 - 1.2223 = -0.4854 \end{aligned}$$

Since $SIG(a, B, D) > SIG(b, B, D)$, the feature a is added to B .

3.3 Feature selection

After computing SIG of features, the final features are selected by a greedy algorithm. In a similar way to [Wang et al., 2017], Algorithm 1 is used to select the final features. Table 2 describes the parameters and variables of the algorithm. In line 1, $A = B$ and none of the features has been selected. In lines 3 to 6, SIG of each feature $a_i \in B$ is computed. In line 7, the feature with the maximum SIG is selected. In line 8, it is checked whether the maximum SIG is greater than the threshold α or not; if it is not, the main loop in lines 2 to 14 is terminated by $start = 0$. Otherwise, the sets B and $SelectedFeatures$ are updated in lines 9 to 10 and this process continues.

Algorithm 1 The greedy algorithm for feature selection

Input: S, A, D, σ, α

Output: $SelectedFeatures$

```

1: Initialize:  $SelectedFeatures \leftarrow \emptyset, B \leftarrow A - SelectedFeatures, start \leftarrow 1;$ 
2: while  $start$  do
3:   for each  $(a_i \in B)$  do
4:     compute the neighborhood relation  $R_{SelectedFeatures \cup \{a_i\}}^p$ ;
5:     compute  $SIG(a_i, SelectedFeatures, D)$ ;
6:   end for
7:   Find  $a_k$  with the maximum  $SIG(a_k, SelectedFeatures, D)$ ;
8:   if  $(SIG(a_k, SelectedFeatures, D) > \alpha)$  then
9:      $SelectedFeatures \leftarrow SelectedFeatures \cup \{a_k\}$ 
10:     $B \leftarrow B - SelectedFeatures$ 
11:   else
12:      $start \leftarrow 0;$ 
13:   end if
14: end while
15: return  $SelectedFeatures;$ 

```

Example 5. In this example, EPSTEIN is employed on the IRIS data set selected from the UCI Machine Learning Repository [Blake, 1998]. There are three species of IRIS in the data set, each species has 50 samples and each sample is described by four features. Assume $\sigma = 0.01$ and $\alpha = 0.001$. The process of the feature selection according to Algorithm 1 is as follows (note that SIG_i is SIG of the feature i):

- **Iteration 1:** $A = B = \{1, 2, 3, 4\}$, $SelectedFeatures = \emptyset$, $SIG_1 = 0$, $SIG_2 = 0.10591$, $SIG_3 = 0.23339$, $SIG_4 = 0$
- **Iteration 2:** $B = \{1, 2, 4\}$, $SelectedFeatures = \{3\}$, $SIG_1 = -0.231334$, $SIG_2 = 0.021879$, $SIG_4 = -0.23339$
- **Iteration 3:** $B = \{1, 4\}$, $SelectedFeatures = \{2, 3\}$, $SIG_1 = -0.00454$, $SIG_4 = 0$

In each iteration, the feature with the maximum $SIG > \alpha$ is selected. In iteration 3, since SIG of all the unselected features (B) is less than α , the algorithm is terminated.

4 Evaluation

In this section, we provide a comprehensive evaluation of EPSTEIN. The classification accuracy of EPSTEIN and the effect of the most important parameters on the feature

selection are considered in this section. We evaluate EPSTEIN on fifteen data sets selected from the UCI Machine Learning Repository [Blake, 1998]. Table 3 shows these data sets. There are two parameters for EPSTEIN: σ and α . The parameters setting for the evaluation of EPSTEIN is as follows:

- α : the small values of α might lead to increasing the number of the selected features, which could increase the duration of the training phase of classifiers and decrease the classification accuracy. On the other hand, the large values of α might lead to the inability to identify all the significant features. In [Wang et al., 2017], the impact of α has been considered and $\alpha = 0.001$ has been selected. So, we also set $\alpha = 0.001$ for evaluation.
- σ : The local density of data points is computed based on σ . The value of σ should be selected in a way that an appropriate subset of data points is used to compute the local density. The small values of σ cause a few points to be considered as the nearest neighbors. On the other hand, the large values of σ cause many points to be considered as the nearest neighbors. Therefore, different values of σ can lead to different classification accuracy; The impact of σ on the classification accuracy and the number of the selected features are evaluated.

EPSTEIN is compared with some state-of-the-art methods such as HANDI [Wang et al., 2017], KNFI [Thejas et al., 2019], KNFE [Thejas et al., 2019], CFS [Hall, 2000], FCBF [Yu and Liu, 2003] and PCA-IG [Omuya et al., 2021]. These methods have been reviewed in section 2.1. Since the parameters have a significant effect on the prediction results of the classifiers [Amiri et al., 2018a, Amiri et al., 2018b, Amiri and Askari, 2022], each method is evaluated by using the best values of its parameters:

- **FCBF**: The relevance threshold of FCBF is set to 0 [Cortes and Vapnik, 1995, Altman, 1992].
- **HANDI**: The threshold δ (a similar parameter to α) of HANDI is set to 0.001 [Wang et al., 2017]. Since the core of EPSTEIN and HANDI are similar, we also investigate the impact of the neighborhood radius (ϵ) on the number of the selected features and the classification accuracy in a similar way to σ .
- **PCA-IG**: In this model, a set threshold t is used to select features based on IG. According to [Prasetyowati et al., 2021], t is set to 0.05.

Three common classifiers RBF-SVM [Cortes and Vapnik, 1995], KNN [Altman, 1992] and Decision Tree [Loh, 2011] are used to evaluate these algorithms. Because our goal is to compare the performance of the different feature selection algorithms, we don't focus on the parameter setting of the classifiers. According to the results reported in [Ho and Wechsler, 2008], the control parameter C is set to 100 and the Gaussian kernel parameter γ is set to 1. We set $K = 7$ for KNN. Table 5 lists the values and parameters of the methods. We employ 5-fold cross validation. So the data sets are randomly divided into five subsets; one is used for testing and the remaining four are used for training. The Feature selection algorithms are employed on the training set; the reduced training and testing sets are classified. After 5 rounds, the average value of the classification accuracy is reported as the final performance. All the algorithms have been implemented in Python and all the experiments run on a machine with an Intel(R) Core(TM) i5 CPU M 480 @ 2.67GHz processor and 16 GB of RAM. The impact of the most important parameters on EPSTEIN and HANDI and the impact of the selected features on the classification accuracy are considered in sections 4.1 and 4.2 respectively. In section 4.3, the time complexity of EPSTEIN and the other methods are considered.

Table 3: The description of data sets [Blake, 1998]

No	Data set	Sample	Features	Class
1	Credit	690	14	2
2	SCADI	70	205	7
3	Forest types	325	27	4
4	Sonar	208	60	2
5	ILPD	583	10	2
6	Spect heart	267	22	2
7	Leaf	340	15	30
8	Thoracic Surgery	470	17	2
9	Seeds	210	7	3
10	Wine	178	13	3
11	IRIS	150	4	3
12	Wpbc	198	33	2
13	Lung cancer	31	56	3
14	Primary tumor	339	17	22
15	Breast tissue	106	9	6

Table 4: The best value of σ for the first ten data sets

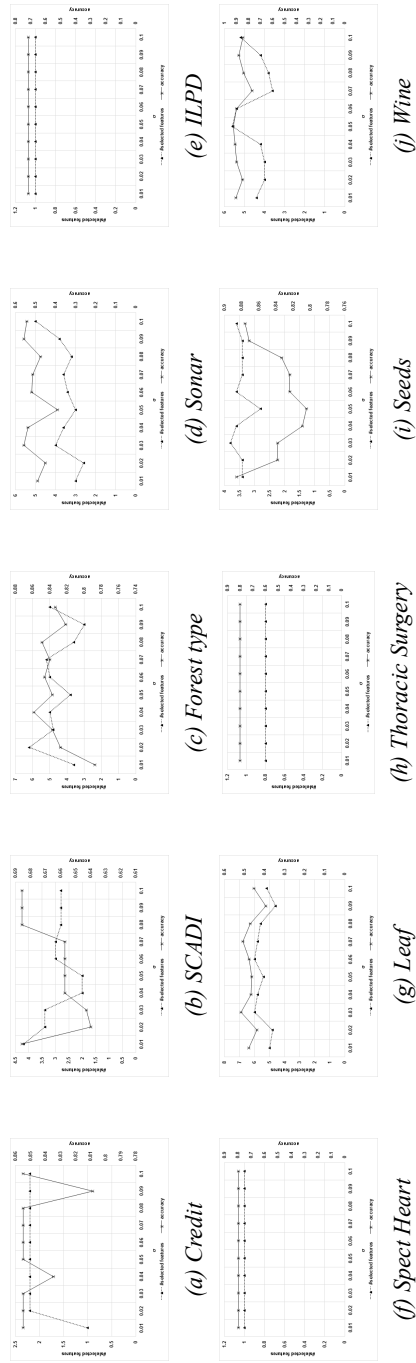
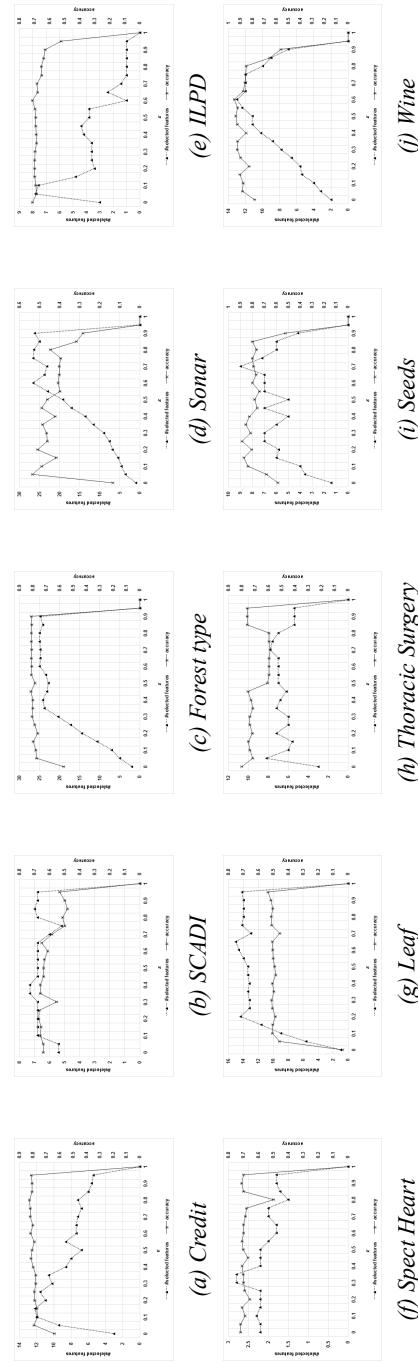
No	Data set	σ
1	Credit	0.01
2	SCADI	0.1
3	Forest types	0.04
4	Sonar	0.03
5	ILPD	[0.01, 0.1] (in the experiments: $\sigma = 0.01$)
6	Spect heart	[0.01, 0.1] (in the experiments: $\sigma = 0.01$)
7	Leaf	0.03
8	Thoracic Surgery	[0.01, 0.1] (in the experiments: $\sigma = 0.01$)
9	Seeds	0.01
10	Wine	0.05

Table 5: The parameters setting of Decision Tree, SVM and KNN

Method	Decision Tree	SVM			KNN
Parameter	Criterion for selecting nodes	Kernel	$gamm$	C	Number of neighbors
Value	Gini Index	RBF	1	100	7

Table 6: The impact of the parameters of EPSTEIN and HANDI (σ and ϵ) on accuracy and selected features

Data Set	Accuracy				Selected Features			
	EPSTEIN (σ)		HANDI (ϵ)		EPSTEIN (σ)		HANDI (ϵ)	
Credit	Min=0.8086	Max=0.855	Min=0	Max=0.8246	Min=1	Max=2.2	Min=0	Max=12.2
	Var=0.0002	Avg=0.8483	Var=0.0312	Avg=0.7528	Var=0.144	Avg=2.08	Var=8.9925	Avg=7.8571
SCADI	Min=0.64	Max=0.6857	Min=0	Max=0.6714	Min=2	Max=4.2	Min=0	Max=7.3
	Var=0.0003	Avg=0.6654	Var=0.0216	Avg=0.5769	Var=0.4271	Avg=2.94	Var=2.4172	Avg=6.2857
Forest types	Min=0.7876	Max=0.8584	Min=0	Max=0.8153	Min=3	Max=6.2	Min=0	Max=25
	Var=0.0003	Avg=0.8337	Var=0.0588	Avg=0.7125	Var=0.9528	Avg=4.52	Var=86.9224	Avg=17.5047
Sonar	Min=0.39	Max=0.5581	Min=0	Max=0.5347	Min=2.6	Max=5	Min=0	Max=26.6
	Var=0.0028	Avg=0.5036	Var=0.0233	Avg=0.3771	Var=0.4462	Avg=3.52	Var=99.6104	Avg=14.2952
ILPD	Min=0.7134	Max=0.7134	Min=0	Max=0.7134	Min=1	Max=1	Min=0	Max=7.8
	Var=0	Avg=0.7134	Var=0.0235	Avg=0.6440	Var=0	Avg=1	Var=4.4636	Avg=3.0190
Spect heart	Min=0.7952	Max=0.7952	Min=0	Max=0.7197	Min=1	Max=1	Min=0	Max=2.8
	Var=0	Avg=0.7952	Var=0.0246	Avg=0.6551	Var=0	Avg=1	Var=0.3094	Avg=1.9952
Leaf	Min=0.3941	Max=0.5176	Min=0	Max=0.5352	Min=4.6	Max=6	Min=0	Max=15
	Var=0.0012	Avg=0.4670	Var=0.0213	Avg=0.4539	Var=0.2528	Avg=5.42	Var=18.6451	Avg=11.8285
Thoracic Surgery	Min=0.8	Max=0.8	Min=0	Max=0.8	Min=0.8	Max=0.8	Min=0	Max=8.2
	Var=0	Avg=0.8	Var=0.0218	Avg=0.6592	Var=0	Avg=0.8	Var=3.2533	Avg=6.1333
Seeds	Min=0.8047	Max=0.8857	Min=0	Max=0.8857	Min=2.8	Max=3.8	Min=0	Max=9
	Var=0.0007	Avg=0.8404	Var=0.0615	Avg=0.6999	Var=0.0693	Avg=3.44	Var=5.7104	Avg=5.2952
Wine	Min=0.7693	Max=0.9269	Min=0	Max=0.9611	Min=3.6	Max=5.6	Min=0	Max=13
	Var=0.0022	Avg=0.8727	Var=0.0751	Avg=0.7749	Var=0.496	Avg=4.44	Var=17.0436	Avg=7.7809
IRIS	Min=0.16	Max=0.9564	Min=0	Max=0.9138	Min=0.2	Max=2.6	Min=0	Max=3.2
	Var=0.0577	Avg=0.5668	Var=0.0587	Avg=0.5753	Var=0.5262	Avg=1.32	Var=0.8702	Avg=2.4857
Wpbc	Min=0.5968	Max=0.9036	Min=0	Max=0.7775	Min=2	Max=3.2	Min=0	Max=3.8
	Var=0.0073	Avg=0.7357	Var=0.0244	Avg=0.6365	Var=0.3004	Avg=2.56	Var=0.7859	Avg=2.3904
Lung cancer	Min=0.1532	Max=0.4125	Min=0	Max=0.3759	Min=2	Max=4.2	Min=0	Max=8.2
	Var=0.0055	Avg=0.2169	Var=0.0058	Avg=0.2823	Var=0.5671	Avg=2.56	Var=3.4871	Avg=5.3285
Primary tumor	Min=0.619	Max=0.7325	Min=0	Max=0.6435	Min=1	Max=1.8	Min=0	Max=13
	Var=0.0012	Avg=0.6310	Var=0.0018	Avg=0.5076	Var=0.1137	Avg=1.64	Var=9.0196	Avg=9.847
Breast tissue	Min=0.1798	Max=0.2315	Min=0	Max=0.2681	Min=2.2	Max=3.4	Min=0	Max=7
	Var=0.0002	Avg=0.1956	Var=0.0032	Avg=0.1612	Var=0.1071	Avg=2.94	Var=2.9942	Avg=5.0142

Figure 4: The impact of σ on the performance of EPSTEIN for the different data setsFigure 5: The impact of ϵ on the performance of HANDI for the different data sets

4.1 Impact of parameters

Since EPSTEIN and HANDI are similar in the basic structure, the impact of σ on EPSTEIN and ϵ on HANDI is evaluated in this section. For this purpose, we consider the average accuracy of the classifier RBF-SVM and the average number of selected features for the first ten data sets. The impact of σ on EPSTEIN in the interval of $[0.01, 0.1]$ with the step of 0.01 and the impact of ϵ on HANDI in the interval of $[0, 1]$ with the step of 0.1 are investigated.

Figure 4 shows the impact of σ on the data sets. According to the figure, in the data sets **Spect heart**, **Thoracic Surgery** and **ILPD**, there is no significant change in both of the accuracy and selected features for different values of σ . In the other data sets, there are minor changes in accuracy and selected features for different values of σ . Table 4 shows the best value of σ , which leads to the maximum value of accuracy. Figure 5 shows the impact of ϵ on the performance of HANDI for the different data sets. According to the figure, for all the data sets, there are significant changes in accuracy and selected features for different values of ϵ . For a comprehensive comparison, table 6 compares the impact of σ on EPSTEIN with ϵ 's on HANDI on all the data sets in terms of *average*, *variance*, *min* and *max* of the accuracy and selected features. As the table shows, the parameter ϵ has a significant influence on the performance of the HANDI algorithm. So selecting an appropriate value for ϵ is very challenging. On the contrary, the impact of σ on the performance of EPSTEIN is more tolerable.

4.2 The impact of the selected features on the classification accuracy

In this section, the impact of the features selected by the different algorithms on the classification accuracy and F-measure is investigated. To provide a fair comparison between EPSTEIN and the other algorithms, the best parameters of the algorithms are determined carefully. The classifiers run 5 times and their average accuracy and F-measure are reported as the performance. Tables 7, 8 and 9 and Figure 6 show the average accuracy of Decision Tree, RBF-SVM and KNN on the different data sets respectively. The average F-measure of Decision Tree, RBF-SVM and KNN is also reported in Figure 7. In addition to the feature selection algorithms, the classification accuracy without feature selection (the column *AF*) is also investigated. In the tables, for each data set, the maximum accuracy is indicated in bold type. For algorithms HANDI and EPSTEIN, the best values of ϵ and σ are also reported (since their best values are the same, we only report them for the average accuracy). To summarize the results, the last row of the tables, indicated by $\langle G, E, L \rangle$, shows the number of data sets that the accuracy of EPSTEIN is, respectively, greater than, equal to and less than the other algorithms'.

Table 7 and Figure 6a show the accuracy of Decision Tree. KNFI and KNFE provide the highest classification accuracy in the data sets SCADI, Sonar, Leaf and Primary tumor. FCBF achieves the highest classification accuracy in the data sets Spect heart and Thoracic Surgery. PCA-IG provides the highest classification accuracy only in the data set Leaf. In the other data sets, EPSTEIN provides the most precise results. According to the last row, EPSTEIN provides more reliable results generally. Table 8 and Figure 6b show the accuracy of RBF-SVM. HANDI provides the highest classification accuracy in the data sets Wine and Breast tissue. FCBF and FCS achieve the highest classification accuracy in the data sets Forest types and Thoracic Surgery. KNFI and KNFE achieve the highest classification accuracy in the data sets Leaf and Primary tumor. In the other data sets, EPSTEIN provides the most precise results. According to the last two rows, EPSTEIN provides more reliable results on 9, 13, 13, 13, 10, 11 and 15 data sets compared

Table 7: The classification accuracy of Decision Tree based on the different feature selection algorithms

data set	EPSTEIN	HANDI	CFS	FCBF	AF	KNFI	KNFE	PCA-IG
Credit	0.8550	0.7840	0.7942	0.7333	0.7826	0.8318	0.8507	0.7391
	$\sigma = 0.01$	$\epsilon = 0.8$						
SCADI	0.6857	0.6571	0.7571	0.5714	0.7285	0.7835	0.7857	0.6581
	$\sigma = 0.1$	$\epsilon = 0.2$						
Forest types	0.7846	0.7661	0.7415	0.7415	0.7446	0.76	0.7538	0.7461
	$\sigma = 0.04$	$\epsilon = 0.45$						
Sonar	0.4754	0.5144	0.4760	0.4331	0.4279	0.5370	0.5968	0.3552
	$\sigma = 0.03$	$\epsilon = 0.05$						
ILPD	0.7134	0.6329	0.6416	0.6773	0.6243	0.6894	0.6312	0.6413
	$\sigma = 0.01$	$\epsilon = 0.6$						
Spect heart	0.7952	0.7197	0.7342	0.7952	0.7342	0.7252	0.7738	0.7009
	$\sigma = 0.01$	$\epsilon = 0.05$						
Leaf	0.3735	0.3882	0.3470	0.2352	0.3735	0.3917	0.3970	0.4029
	$\sigma = 0.03$	$\epsilon = 0.95$						
Thoracic Surgery	0.8	0.6319	0.7531	0.8446	0.7361	0.7831	0.7940	0.5129
	$\sigma = 0.01$	$\epsilon = 0$						
Seeds	0.7904	0.7809	0.7761	0.7523	0.7571	0.7619	0.7238	0.7666
	$\sigma = 0.01$	$\epsilon = 0.25$						
Wine	0.8712	0.8480	0.7917	0.4871	0.8257	0.8538	0.8198	0.7902
	$\sigma = 0.05$	$\epsilon = 0.6$						
IRIS	0.9645	0.9228	0.9133	0.5	0.9124	0.9533	0.9533	0.8933
	$\sigma = 0.01$	$\epsilon = 0$						
Wpbc	0.8953	0.8831	0.5851	0.8505	0.7721	0.6714	0.6252	0.7017
	$\sigma = 0.04$	$\epsilon = 0.45$						
Lung cancer	0.3911	0.3759	0.2857	0.1190	0.3521	0.2034	0.2321	0.2142
	$\sigma = 0.02$	$\epsilon = 0.5$						
Primary tumor	0.6137	0.6436	0.6345	0.6343	0.6258	0.7287	0.6816	0.6346
	$\sigma = 0.01$	$\epsilon = 0.95$						
Breast tissue	0.2795	0.2681	0.2281	0.1809	0.2169	0.2663	0.2757	0.2042
	$\sigma = 0.04$	$\epsilon = 0.3$						
Average	0.6841	0.6544	0.6311	0.5638	0.6395	0.6627	0.6596	0.5974
< G, E, L >	—	< 12, 0, 3 >	< 11, 0, 4 >	< 12, 1, 2 >	< 12, 1, 1 >	< 11, 0, 4 >	< 11, 0, 4 >	< 13, 0, 2 >

Table 8: The classification accuracy of RBF-SVM based on the different feature selection algorithms

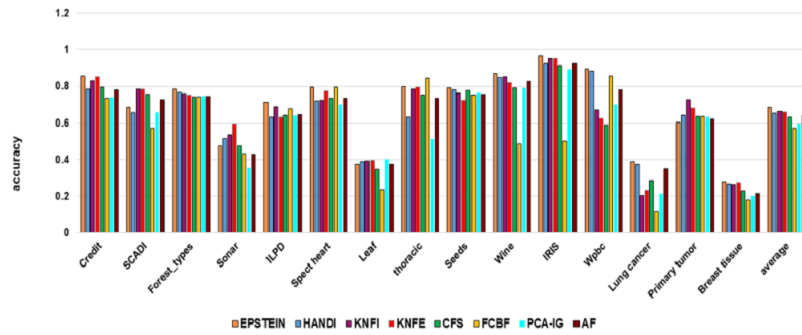
data set	EPSTEIN	HANDI	CFS	FCBF	AF	KNFI	KNFE	PCA-IG
Credit	0.8550	0.8246	0.8188	0.7333	0.7811	0.8333	0.8320	0.8057
	$\sigma = 0.01$	$\epsilon = 0.8$						
SCADI	0.6857	0.6714	0.6714	0.6	0.4142	0.5048	0.4142	0.5285
	$\sigma = 0.1$	$\epsilon = 0.2$						
Forest types	0.8584	0.8153	0.8646	0.8430	0.8184	0.7152	0.6895	0.84
	$\sigma = 0.04$	$\epsilon = 0.45$						
Sonar	0.5581	0.5347	0.4376	0.1975	0.3209	0.5501	0.5428	0.3974
	$\sigma = 0.03$	$\epsilon = 0.05$						
ILPD	0.7134	0.7134	0.7100	0.7134	0.6809	0.7134	0.7083	0.7048
	$\sigma = 0.01$	$\epsilon = 0.6$						
Spect heart	0.7952	0.7197	0.7944	0.7952	0.7895	0.7952	0.7809	0.6828
	$\sigma = 0.01$	$\epsilon = 0.05$						
Leaf	0.5176	0.5352	0.4970	0.1792	0.5352	0.5264	0.6117	0.4852
	$\sigma = 0.03$	$\epsilon = 0.95$						
Thoracic Surgery	0.8	0.8	0.8042	0.8489	0.7893	0.8510	0.8297	0.6147
	$\sigma = 0.01$	$\epsilon = 0$						
Seeds	0.8857	0.8857	0.8666	0.7666	0.8857	0.8428	0.8238	0.8285
	$\sigma = 0.01$	$\epsilon = 0.25$						
Wine	0.9269	0.9611	0.9158	0.4985	0.9496	0.9247	0.8938	0.8823
	$\sigma = 0.05$	$\epsilon = 0.6$						
IRIS	0.9564	0.9138	0.9266	0.4466	0.9102	0.9466	0.9466	0.92
	$\sigma = 0.01$	$\epsilon = 0$						
Wpbc	0.9036	0.7775	0.6664	0.8946	0.8821	0.7210	0.7316	0.7061
	$\sigma = 0.04$	$\epsilon = 0.45$						
Lung cancer	0.4125	0.3759	0.3095	0.090	0.3529	0.2158	0.1849	0.1619
	$\sigma = 0.02$	$\epsilon = 0.5$						
Primary tumor	0.7325	0.6436	0.6845	0.6607	0.7280	0.7287	0.7405	0.6280
	$\sigma = 0.01$	$\epsilon = 0.95$						
Breast tissue	0.2315	0.2681	0.2186	0.1142	0.2756	0.2	0.26	0.2059
	$\sigma = 0.04$	$\epsilon = 0.3$						
Average	0.7219	0.696	0.6790	0.5587	0.6755	0.6712	0.6660	0.6261
< G, E, L >	—	< 9, 2, 4 >	< 13, 0, 2 >	< 13, 2, 0 >	< 13, 1, 1 >	< 10, 2, 3 >	< 11, 0, 4 >	< 15, 0, 0 >

Table 9: The classification accuracy of 7NN based on the different feature selection algorithms

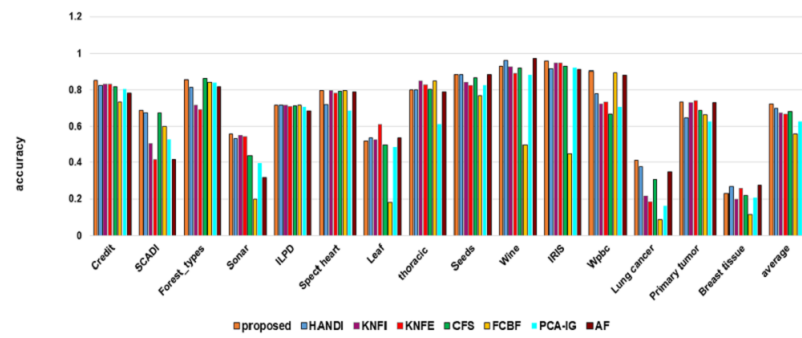
data set	EPSTEIN	HANDI	CFS	FCBF	AF	KNFI	KNFE	PCA-IG
Credit	0.8275	0.7449	0.8217	0.4202	0.8104	0.8057	0.7753	0.8031
	$\sigma = 0.01$	$\epsilon = 0.8$						
SCADI	0.7442	0.7652	0.7371	0.6857	0.7124	0.7714	0.7857	0.7428
	$\sigma = 0.1$	$\epsilon = 0.2$						
Forest types	0.8276	0.8030	0.8153	0.7846	0.8122	0.8038	0.8030	0.7538
	$\sigma = 0.04$	$\epsilon = 0.45$						
Sonar	0.5232	0.5439	0.4149	0.3081	0.3682	0.6825	0.5473	0.3931
	$\sigma = 0.03$	$\epsilon = 0.05$						
ILPD	0.6998	0.6774	0.6534	0.6637	0.6137	0.6791	0.6604	0.6620
	$\sigma = 0.01$	$\epsilon = 0.6$						
Spect heart	0.7205	0.6713	0.7619	0.7529	0.7603	0.7730	0.7327	0.6635
	$\sigma = 0.01$	$\epsilon = 0.05$						
Leaf	0.3522	0.3117	0.3323	0.1617	0.3418	0.3041	0.3152	0.3235
	$\sigma = 0.03$	$\epsilon = 0.95$						
Thoracic Surgery	0.8	0.7510	0.8106	0.8268	0.8197	0.7910	0.8271	0.7031
	$\sigma = 0.01$	$\epsilon = 0.25$						
Seeds	0.8619	0.8476	0.8571	0.7809	0.8325	0.7428	0.8095	0.8238
	$\sigma = 0.01$	$\epsilon = 0.25$						
Wine	0.9158	0.9041	0.8988	0.5488	0.9104	0.9028	0.8993	0.8542
	$\sigma = 0.05$	$\epsilon = 0.6$						
IRIS	0.9189	0.9237	0.9133	0.4533	0.9045	0.96	0.9533	0.94
	$\sigma = 0.01$	$\epsilon = 0$						
Wpbc	0.8921	0.8647	0.7420	0.8875	0.7524	0.7619	0.7366	0.6607
	$\sigma = 0.04$	$\epsilon = 0.45$						
Lung cancer	0.38	0.3781	0.1904	0.0571	0.2512	0.2218	0.2129	0.0285
	$\sigma = 0.02$	$\epsilon = 0.5$						
Primary tumor	0.7333	0.71	0.6445	0.4549	0.7044	0.6578	0.7199	0.6022
	$\sigma = 0.01$	$\epsilon = 0.95$						
Breast tissue	0.181	0.2599	0.0852	0.019	0.21	0.0190	0.0852	0.2042
	$\sigma = 0.04$	$\epsilon = 0.3$						
Average	0.6918	0.6771	0.6452	0.5203	0.6536	0.6584	0.6575	0.6105
$\langle G, E, L \rangle$	—	$\langle 11, 0, 4 \rangle$	$\langle 12, 0, 3 \rangle$	$\langle 13, 0, 2 \rangle$	$\langle 12, 0, 3 \rangle$	$\langle 11, 0, 4 \rangle$	$\langle 11, 0, 4 \rangle$	$\langle 13, 0, 2 \rangle$

Table 10: The average number of the features selected by the different algorithms

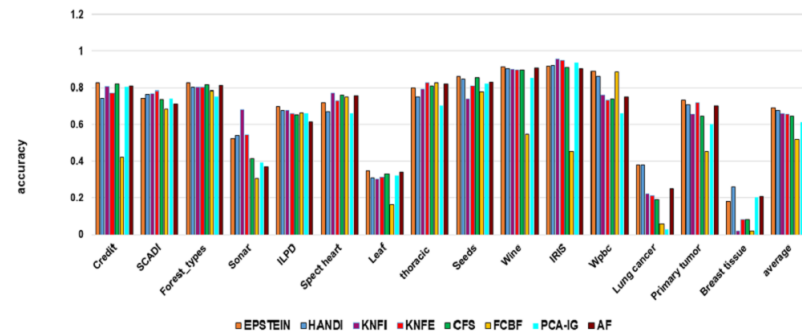
data set	EPSTEIN	HANDI	CFS	FCBF	AF	KNFI	KNFE	PCA-IG
Credit	1	7.2	6	2	14	8.6	3.2	3
SCADI	2.8	6.8	23.4	11.2	205	3	199.4	16
Forest types	5	3.2	9.8	5.2	27	3.2	7	5.8
Sonar	4	3.6	6.2	2	60	4.2	5.4	6
ILPD	1	1	6.2	1	10	1.8	7.8	2
Spect Heart	1	2.2	15.6	1	22	1	13.2	4
Leaf	6	14.2	7.2	2	15	6.6	13.4	4.4
Thoracic surgery	0.8	7	8	1	17	1.2	8.8	1
Seeds	3.4	3	6.6	1.8	7	4.4	4.4	2.8
Wine	5.6	13	6.4	1	13	4	9.6	7.2
IRIS	2.6	2.8	6	2	4	1.2	2.8	3
Wpbc	2	1.8	6.8	2	33	4.8	17.2	2.4
Lung cancer	4.2	4	15.6	1.8	56	2	2.3	10
Primary tumor	1	7.8	9.6	1	17	9.2	10.4	3.4
Breast tissue	3	4.8	7.6	2	9	2	7.2	2.6



(a) Decision Tree

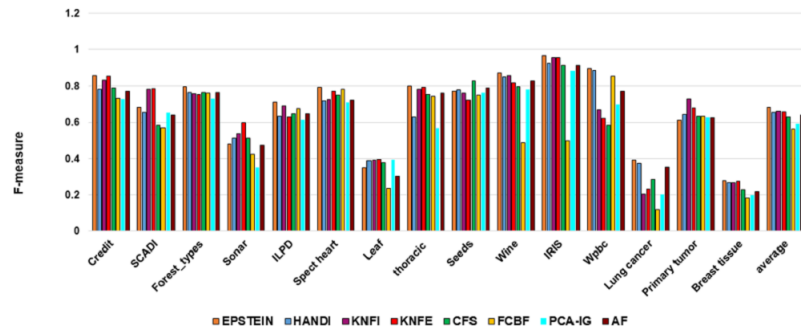


(b) RBF-SVM

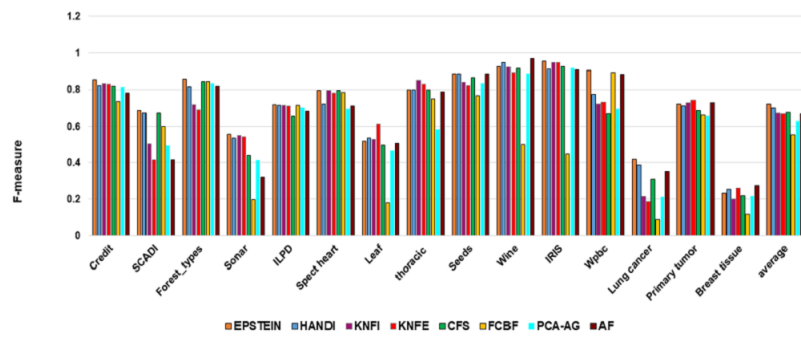


(c) 7NN

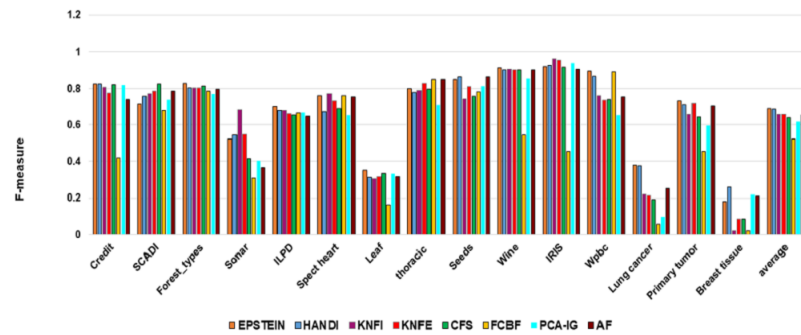
Figure 6: The average accuracy of the classifiers based on the different feature selection algorithms for all the data sets



(a) Decision Tree



(b) RBF-SVM



(c) 7NN

Figure 7: The average F-measure of the classifiers based on the different feature selection algorithms for all the data sets

to HANDI, CFS, FCBF, AF, KNFI, KNFE and PCA-IG respectively. In none of the data sets, PCA-IG could not provide reliable results compared to HANDI and EPSTEIN. Table 9 and Figure 6c show the accuracy of 7NN. HANDI provides the highest classification accuracy in the data set Breast tissue. KNFI and KNFE achieve the highest classification accuracy in the data sets SCADI, Sonar, Spect heart, Thoracic Surgery and IRIS. In the other data sets, EPSTEIN provides the most precise results. According to the last two rows, EPSTEIN provides more reliable results on 11, 12, 13, 12, 11, 11 and 13 data sets compared to HANDI, CFS, FCBF, AF, KNFI, KNFE and PCA-IG respectively. PCA-IG achieves more classification accuracy than EPSTEIN in the data sets IRIS and Breast tissue. Figure 7 shows the F-measure of Decision Tree, RBF-SVM and 7NN. According to the tables and the figure, for all the classifiers, EPSTEIN improves the F-measure compared to the other algorithms on most of the data sets. As the last two rows of the tables show, EPSTEIN provides more reliable results on average compared to the other algorithms. Table 10 shows the average number of the features selected by the different algorithms for each data set. As the results show, EPSTEIN selects fewer features compared to the other algorithms on average. So not only does EPSTEIN alleviate the impact of the neighborhood radius on the neighborhood relation, but it also improves the classification accuracy and decreases the number of the selected features, which can lead to decreasing time complexity of classification.

4.3 Time Complexity Analysis

As our experimental results in section 4.2 show, EPSTEIN, HANDI, KNFI and KNFE provide the most reliable results. So, due to space limitation, we compare the time complexity of EPSTEIN with HANDI's, KNFI's and KNFE's briefly:

- KNFI and KNFE: Let n be the number of samples in each batch of mini-batch Kmeans clustering, m be the number of the features, k be the number of clusters and t be the number of iterations of mini-batch Kmeans. The time complexity of the filter method is $O(n \times m \times k \times t)$. The time complexity of the wrapper method depends on the classifier used to obtain the subset of features. If the time complexity of the classifier is $O(y)$, then the time complexity of the wrapper method is $O(m \times y)$.
- HANDI: Let n be the number of samples and m be the number of the features. The time complexity of neighborhood relations construction is $O(n^2)$. The overall time complexity of the algorithm is $O(n^2 \times m)$.
- EPSTEIN: Let n be the number of samples and m be the number of the features. The time complexity of neighborhood relations construction is $O(n^3)$. The overall time complexity of the algorithm is $O(n^3 \times m)$.

As time complexities and the experimental results show, EPSTEIN improves the other methods with comparable complexity.

5 Conclusion and Future Work

In this paper, we present EPSTEIN, a new approach to construct the neighborhood relation. In EPSTEIN, the target points are determined based on their local density. Then, the circles centered at the target points are drawn and the points inside or on

the circles are considered to be neighbors. In the next steps, based on the conditional discrimination index and the greedy algorithm, an appropriate subset of features is selected. The performance of EPSTEIN is compared to HANDI, KNFE, KNFI, CFS and FCBF on fifteen data sets. According to the experiment results, 1) EPSTEIN alleviates the impact of the neighborhood radius on the neighborhood relation, 2) EPSTEIN improves the classification accuracy and 3) EPSTEIN decreases the number of the selected features, which can lead to decreasing time complexity of classification.

In the future work, we plan to conduct more experiments to evaluate the efficiency of EPSTEIN on larger data sets. We also plan to consider outliers and missing values. Furthermore, we plan to propose an approach to select the values of the parameters according to the characteristic of data sets.

References

- [Aghdam et al., 2009] Aghdam, M. H., Ghasem-Aghaee, N., and Basiri, M. E.: "Text feature selection using ant colony optimization"; *Exp. Sys. App.*, 36, 2 (2009), 6843–6853.
- [Altman, 1992] Altman, N. S.: "An introduction to kernel and nearest-neighbor nonparametric regression"; *Amer. Stat.*, 46, 3 (1992), 175–185.
- [Amiri et al., 2018a] Amiri, M., Mohammad-Khanli, L., and Mirandola, R.: "An online learning model based on episode mining for workload prediction in cloud"; *Fut. Gen. Comp. Sys.*, 87 (2018a), 83–101.
- [Amiri et al., 2018b] Amiri, M., Mohammad-Khanli, L., and Mirandola, R.: "A sequential pattern mining model for application workload prediction in cloud environment"; *Jrnl. Net. Comp. App.*, 105 (2018b), 21–62.
- [Amiri et al., 2020] Amiri, M., Mohammad-Khanli, L., and Mirandola, R.: "A new efficient approach for extracting the closed episodes for workload prediction in cloud"; *Computing*, 102, 1 (2020), 141–200.
- [Amiri and Askari, 2022] Amiri, M., Askari, H.: "Illegal Miner Detection based on Pattern Mining: A Practical Approach"; *Jrnl. Com. Sec.*, 2, 9 (2022), 1–10.
- [Armanfard et al., 2015] Armanfard, N., Reilly, J. P., and Komeili, M.: "Local feature selection for data classification"; *IEEE Trans. Pat. Anal. & Mach. Intell.*, 38, 6 (2015), 1217–1227.
- [Battiti, 1994] Battiti, R.: "Using mutual information for selecting features in supervised neural net learning"; *IEEE Trans. NN.*, 5, 4 (1994), 537–550.
- [Blake, 1998] Blake, C.: "UCI repository of machine learning databases"; (1998) <http://www.ics.uci.edu/~mlearn/MLRepository.html>
- [Bonabeau et al., 1999] Bonabeau, E., Dorigo, M., Marco, D. d. R. D. F., Theraulaz, G., Theraulaz, G., et al.: "Swarm intelligence: from natural to artificial systems"; No. 1, Oxford U. press (1999).
- [Bouaguel, 2016] Bouaguel, W.: "A new approach for wrapper feature selection using genetic algorithm for big data"; *Intell. & Evol. Sys.*, Springer (2016), 75–83.
- [Chen and Chen, 2015] Chen, G. and Chen, J.: "A novel wrapper method for feature selection and its applications"; *Neurocomputing*, 159 (2015), 219–226.
- [Cicioğlu, 2021] Cicioğlu, M.: "Multi-criteria handover management using entropy-based saw method for sdn-based 5g small cells"; *Wirel. Net.*, 27, 4 (2021), 2947–2959.
- [Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V.: "Support-vector networks"; *ML.*, 20, 3 (1995), 273–297.
- [Dai et al., 2012] Dai, J., Wang, W., and Xu, Q.: "An uncertainty measure for incomplete decision tables and its applications"; *IEEE Trans. Cybern.*, 43, 4 (2012), 1277–1289.

- [Dong and Liu, 2018] Dong, G. and Liu, H.: “Feature engineering for machine learning and data analytics”; CRC Press (2018).
- [Du et al., 2016] Du, M., Ding, S., and Jia, H.: “Study on density peaks clustering based on k-nearest neighbors and principal component analysis”; *Knowl. Based Sys.*, 99 (2016), 135–145.
- [Fard et al., 2013] Fard, S. M. H., Hamzeh, A., and Hashemi, S.: “Using reinforcement learning to find an optimal set of features”; *Comp. & Math. App.*, 66, 10 (2013), 1892–1904.
- [Gaudel and Sebag, 2010] Gaudel, R. and Sebag, M.: “Feature selection as a one-player game”; *Internat. Conf. ML.* (2010), 359–366.
- [Goldberg, 1989] Goldberg, D. E.: “Genetic algorithms in search”; *Optim., & ML.* (1989).
- [Hall, 2000] Hall, M. A.: “Correlation-based feature selection for discrete and numeric class machine learning”; *Proc. 17th Internat. Conf. ML.* (2000), 359–366.
- [Hasanloei et al., 2018] Hasanloei, M. A. V., Sheikhpour, R., Sarram, M. A., Sheikhpour, E., and Sharifi, H.: “A combined fisher and laplacian score for feature selection in qsar based drug design using compounds with known and unknown activities”; *Jrnl. Comput. Aided Mol. Des.*, 32, 2 (2018), 375–384.
- [Ho and Wechsler, 2008] Ho, S.-S. and Wechsler, H.: “Query by transduction”; *IEEE Trans. Pat. Anal. Mach. Intell.*, 30, 9 (2008), 1557–1571.
- [Hu et al., 2008a] Hu, Q., Yu, D., Liu, J., and Wu, C.: “Neighborhood rough set based heterogeneous feature subset selection”; *Info. Sci.*, 178, 18 (2008a), 3577–3594.
- [Hu et al., 2008b] Hu, Q., Yu, D., Liu, J., and Wu, C.: “Neighborhood rough set based heterogeneous feature subset selection”; *Info. Sci.*, 178, 18 (2008b), 3577–3594.
- [Hu et al., 2011] Hu, Q., Zhang, L., Zhang, D., Pan, W., An, S., and Pedrycz, W.: “Measuring relevance between discrete and continuous features based on neighborhood mutual information”; *Exp. Sys. App.*, 38, 9 (2011), 10737–10750.
- [Jiang and Wang, 2016] Jiang, S.-y. and Wang, L.-x.: “Efficient feature selection based on correlation measure between continuous and discrete features”; *Info. Presg. Lett.*, 116, 2 (2016), 203–215.
- [Kamalov and Thabtah, 2017] Kamalov, F. and Thabtah, F.: “A feature selection method based on ranked vector scores of features for classification”; *Ann. Data Sci.*, 4, 4 (2017), 483–502.
- [Kira and Rendell, 1992] Kira, K. and Rendell, L. A.: “A practical approach to feature selection”; *ML. Proc. 1992, Elsevier* (1992), 249–256.
- [Kononenko, 1994] Kononenko, I.: “Estimating attributes: analysis and extensions of relief”; *Eur. Conf. ML., Springer* (1994), 171–182.
- [Li et al., 2013] Li, H., Zhou, X., Zhao, J., and Liu, D.: “Non-monotonic attribute reduction in decision-theoretic rough sets”; *Fundam. Inform.*, 126, 4 (2013), 415–432.
- [Li and Tang 2018] Li, Z. and Tang, Y.: “Comparative density peaks clustering”; *Exp. Sys. App.*, 95 (2018), 236–247.
- [Liu et al., 2017] Liu, J., Lin, Y., Lin, M., Wu, S., and Zhang, J.: “Feature selection based on quality of information”; *Neurocomputing*, 225 (2017), 11–22.
- [Liu et al., 2019] Liu, K., Fu, Y., Wang, P., Wu, L., Bo, R., and Li, X.: “Automating feature subspace exploration via multi-agent reinforcement learning”; *Proc. 25th ACM SIGKDD Internat. Conf. Knowl. Discov. & DM.* (2019), 207–215.
- [Loh, 2011] Loh, W.-Y.: “Classification and regression trees”; *Wiley Interdiscip. Rev.: DM. & Knowl. Discov.*, 1, 1 (2011), 14–23.
- [Mafarja and Mirjalili, 2018] Mafarja, M. and Mirjalili, S.: “Whale optimization approaches for wrapper feature selection”; *Appl. Soft Comput.*, 62 (2018), 441–453.

- [Mariello and Battiti, 2018] Mariello, A. and Battiti, R.: "Feature selection based on the neighborhood entropy"; IEEE Trans. NN. Learning Sys., 29, 12 (2018), 6313–6322.
- [Omuya et al., 2021] Omuya, E. O., Okeyo, G. O., and Kimwele, M. W.: "Feature selection for classification using principal component analysis and information gain"; Exp. Sys. App., 174 (2021), 114765.
- [Peng et al., 2005] Peng, H., Long, F., and Ding, C.: "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy"; IEEE Trans. Pat. Anal. Mach. Intell., 27, 8 (2005), 1226–1238.
- [Pourbahrami et al., 2019] Pourbahrami, S., Khanli, L. M., and Azimpour, S.: "A novel and efficient data point neighborhood construction algorithm based on apollonius circle"; Exp. Sys. App., 115 (2019), 57–67.
- [Prasetyowati et al., 2021] Prasetyowati, M. I., Maulidevi, N. U., and Surendro, K.: "Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest"; Jnl. Big Data, 8, 1 (2021), 1–22.
- [Pudil et al., 1994] Pudil, P., Novovičová, J., and Kittler, J.: "Floating search methods in feature selection"; Pat. Recog. Lett., 15, 11 (1994), 1119–1125.
- [Sahebi et al., 2020] Sahebi, G., Movahedi, P., Ebrahimi, M., Pahikkala, T., Plosila, J., and Tenhunen, H.: "Gefes: A generalized wrapper feature selection approach for optimizing classification performance"; Comp. Boil. Med., 125 (2020), 103974.
- [Shannon, 2001] Shannon, C. E.: "A mathematical theory of communication"; ACM SIGMOBILE mobile Comput. Comms. Rev., 5, 1 (2001), 3–55.
- [Sun et al., 2019a] Sun, L., Zhang, X., Qian, Y., Xu, J., and Zhang, S.: "Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification"; Info. Sci., 502 (2019a), 18–41.
- [Sun et al., 2019b] Sun, L., Zhang, X.-Y., Qian, Y.-H., Xu, J.-C., Zhang, S.-G., and Tian, Y.: "Joint neighborhood entropy-based gene selection method with fisher score for tumor classification"; Appl. Intell., 49, 4 (2019b), 1245–1259.
- [Suresh and Narayanan, 2019] Suresh, S. and Narayanan, A.: "Improving classification accuracy using combined filter+ wrapper feature selection technique"; 2019 IEEE Internat. Conf. Electr., Comp. & Communic. Tech. (ICECCT), IEEE (2019), 1–6.
- [Thejas et al., 2019] Thejas, G., Joshi, S. R., Iyengar, S. S., Sunitha, N., and Badrinath, P.: "Mini-batch normalized mutual information: A hybrid feature selection method"; IEEE Acs., 7 (2019), 116875–116885.
- [Wang et al., 2017] Wang, C., Hu, Q., Wang, X., Chen, D., Qian, Y., and Dong, Z.: "Feature selection based on neighborhood discrimination index"; IEEE Trans. NN learning Sys., 29, 7 (2017), 2986–2999.
- [Wei et al., 2020] Wei, G., Zhao, J., Feng, Y., He, A., and Yu, J.: "A novel hybrid feature selection method based on dynamic feature importance"; Appl. Soft Comput.(2020), 106337.
- [Yang and Ong, 2011] Yang, J.-B. and Ong, C.-J.: "Feature selection using probabilistic prediction of support vector regression"; IEEE Trans. NN., 22, 6 (2011), 954–962.
- [Yu and Liu, 2003] Yu, L. and Liu, H.: "Feature selection for high-dimensional data: A fast correlation-based filter solution"; Proc. 20th Internat. Conf. ML. (ICML-03) (2011), 856–863.
- [Yu and Liu, 2004] Yu, L. and Liu, H.: "Efficient feature selection via analysis of relevance and redundancy"; Jnl. ML. Res., 5 (Oct 2004), 1205–1224.
- [Zhu and Hu, 2013] Zhu, P. and Hu, Q.: "Adaptive neighborhood granularity selection and combination based on margin distribution optimization"; Info. Sci., 249 (2013), 1–12.