Unsupervised Feature Value Selection Based on Explainability

Kilho Shin[1](✉), Kenta Okumoto[2], David Lawrence Shepard[3], Akira Kusaba[6],
Takako Hashimoto[4], Jorge Amari[1], Keisuke Murota[1], Junnosuke Takai[1],
Tetsuji Kuboyama[1], and Hiroaki Ohshima[5]

[1] Gakushuin University, Tokyo, Japan
yoshihiro.shin@gakushuin.ac.jp
[2] Japan Post Bank, Tokyo, Japan
[3] Evidental Health, California, CA, USA
[4] Chiba University of Commerce, Chiba, Japan
[5] University of Hyogo, Kobe, Japan
[6] Kyushu University, Fukuoka, Japan

Abstract. The problem of feature selection has been an area of considerable research in machine learning. Feature selection is known to be particularly difficult in unsupervised learning because different subgroups of features can yield useful insights into the same dataset. In other words, many theoretically-right answers may exist for the same problem. Furthermore, designing algorithms for unsupervised feature selection is technically harder than designing algorithms for supervised feature selection because unsupervised feature selection algorithms cannot be guided by class labels. As a result, previous work attempts to discover intrinsic structures of data with heavy computation such as matrix decomposition, and require significant time to find even a single solution. This paper proposes a novel algorithm, named Explainability-based Unsupervised Feature Value Selection (EUFVS), which enables a paradigm shift in feature selection, and solves all of these problems. EUFVS requires only a few tens of milliseconds for datasets with thousands of features and instances, allowing the generation of a large number of possible solutions and select the solution with the best fit. Another important advantage of EUFVS is that it selects feature values instead of features, which can better explain phenomena in data than features. EUFVS enables a paradigm shift in feature selection. This paper explains its theoretical advantage, and also shows its applications in real experiments. In our experiments with labeled datasets, EUFVS found feature value sets that explain labels, and also detected useful relationships between feature value sets not detectable from given class labels.

Keywords: Feature selection · Unsupervised learning · Clustering

1 Introduction

Feature selection is one of the classical problems of machine learning. While many methods have been developed for supervised learning because the problem is relatively

A. P. Rocha et al. (Eds.): ICAART 2020, LNAI 12613, pp. 421–444, 2021.
https://doi.org/10.1007/978-3-030-71158-0_20

easy given the presence of class labels, in unsupervised learning, no such labels are available and the problem is classically hard.

In supervising learning, a target phenomenon is predetermined and is described in a dataset through class labels associated with individual instances. Each instance of the dataset is a vector of values of the same dimensionality, and each dimension is referred to as a feature. The objective of feature selection in supervised learning is to select as few features as possible with high explanatory ability of the target phenomenon. Since it is theoretically evident that fewer features cannot have more explanatory ability, the objective of supervised feature selection is to find an optimal balance to this trade-off between the number of features and the explanatory ability that they bear. The explanatory ability of features, however, can be understood in multiple ways. One typical way is to define it through statistical or information-theoretic indices like correlation coefficients and mutual information. Another is to define it as the potential predictive power of the features, which can be measured by accuracy of prediction by classifiers, when the values of the features and the class labels of instances are input into the classifiers. Different definitions of explanatory ability may lead us to different conclusions to the question of what is the best result of feature selection. In fact, feature selection methods for obtaining high explanatory ability defined through statistical and information-theoretic indices are categorized as filter-type feature selection, while methods belonging to the wrapper-type and embedded-type feature selection aim to realize high predictive performance for particular classifiers. Nevertheless, it is common in any case that the target phenomenon is given and unchanging, and instances' class labels play a critical role in feature selection.

In contrast, unsupervised feature selection operates without a definite solution or source of truth, because a target phenomenon is not pre-defined. What counts as the "right" result in unsupervised feature selection is unclear. We refer to this as the *indefiniteness problem*. It is as if we were traveling without knowing our destination, and had to decide if we reached the right destination when we arrived.

One possible solution to the indefiniteness problem is using clustering to generate pseudo-labels for each instance. Clustering is the process of categorizing instances based on their similarity. For example, when instances are plotted as points in a Euclidean space, similarity between two instances can be defined as the Euclidean distance between the corresponding points. By assuming that clusters define class labels, we can reduce unsupervised feature selection to supervised feature selection. Eventually, some unsupervised feature selection algorithms proposed in the literature first determine pseudo-labels through clustering and then apply supervised feature selection to explain the pseudo-labels [8,9,12].

Using clustering to generate pseudo-class labels, however, does not solve the indefiniteness problem, because diverse definitions of similarity exist, and different definitions lead to different sets of selected features. For example, the L^∞ distance between $(x_1,\ldots,x_n)$ and $(y_1,\ldots,y_n)$ is identical to $\max\{|x_i - y_i| \mid i = 1,\ldots,n\}$ and evidently yields a totally different similarity measure than the Euclidean distance does. This issue also occurs for other methods known in the literature like methods to select features so as to preserve manifold structures [4,7,25] and data-specific structures [21,22].

In principle, the indefiniteness problem cannot be solved, as shown by the following thought experiment. A DNA array of human beings determines a sequence of genes, and each gene corresponds to an individual biological function. Given a particular biological function, for example, a particular genetic disease, identifying the gene that causes the function is nothing more than supervised feature selection: each gene is a feature. Unsupervised feature selection for a DNA array requires to identify *some* gene without specifying a particular biological function, and we see that, in theory, a great number of right answers exists.

Thus, in this paper, we accept the indefiniteness problem as an inherent limitation of unsupervised feature selection and propose a new approach to address this issue. The key is the development of an unsupervised feature selection algorithm with the following properties:

- High time efficiency
- Hyperparameters for selecting different features

By leveraging such an algorithm, we can run many iterations of the algorithm with different hyperparameter values and can obtain many different answers (sets of features). From these results, we choose the most appropriate answer according to our purpose.

In fact, the main contribution of this paper is to propose a novel algorithm, namely, *Explainability-based Unsupervised Feature Value Selection* (EUFVS), which is both highly efficient and has hyperparameters for selecting different feature sets. In fact, EUFVS requires only a few tens of milliseconds to obtain a single answer for a dataset with thousands of features and thousands of instances. EUFVS is based on the algorithm presented in [19]. For example, EUFVS selects feature values instead of features. This idea was initially introduced in [19] and yields the advantages of more concrete and more efficient interpretation of selection results. On the other hand, EUFVS has two key differences:

- EUFVS is theoretically based on the novel concept of *explainability*;
- EUFVS takes two hyper-parameters rather than a single hyper-parameter, which change the search space of the algorithm in two independent directions, and as a result, can output a wider range of answers.

This paper is organized as follows. Section 2 introduces some mathematical notations and explains some mathematical concepts used in this paper. Section 3 compares supervised and unsupervised feature selection in more detail, and Sect. 4 explains the advantages of feature value selection over feature selection. In Sect. 5, we introduce the concept of explainability and our algorithm, EUFVS. Section 6 is devoted to reporting the results of our experiments to evaluate effectiveness of our algorithm.

2 Formalization and Notations

In this paper, a dataset D is a set of instances, and $\mathcal{F}$ denotes the entire set of the features that describe D. A feature $f \in \mathcal{F}$ is a function $f : D \to R(f)$, where $R(f)$ denotes the range of f, which is a finite set of values.

More formally, we canonically determine a probability space as follows and can view features as random variables. We define a sample space Ω as $\prod_{f\in\mathcal{F}} R(f)$ and a σ-algebra Σ as $\mathfrak{P}(\Omega)$, the power set of Ω. Then, the dataset D introduces an empirical probability measure $p : \mathfrak{P}(\Omega) \to [0,1]$: For an element $\boldsymbol{v} \in \Omega$, $p(\{\boldsymbol{v}\})$ is determined by the ratio of the number of occurrences of $\boldsymbol{v}$ in D to the size of D, that is, $p(\{\boldsymbol{v}\}) = \frac{|\{x\in D|x=\boldsymbol{v}\}|}{|D|}$; For a set $S \in \mathfrak{P}(\Omega)$, we let $p(S) = \sum_{\boldsymbol{v}\in S} p(\{\boldsymbol{v}\})$. Evidently, the triplet (Ω,Σ,p) determines a probability space, which is also known as an *empirical probability space*. Furthermore, we identify a feature f with the projection $\pi_f : \Omega \to R(f)$, and hence, we can view f as a random variable. Thus, we can view f in two different ways, as a function $f : D \to R(f)$ and as a function $f : \prod_{f\in S} R(f) \to R(f)$. This is natural, however, because the support of a multiset D is a subset of Ω.

Moreover, we identify a finite set of features $\{f_1,\ldots,f_n\} \subseteq \mathcal{F}$ with the product of random variables $f_1 \times \cdots \times f_n : \Omega \to R(f_1) \times \cdots \times R(f_n)$. Under this definition, the probability distribution for a feature set is identical to the joint probability distribution of the random variables (features) involved.

By viewing feature sets, say S and T, as random variables, we can apply many useful information theoretical indices to S and T. Such indices include information entropy $H(S)$, mutual information $I(S;T)$, normalized mutual information $\mathrm{NMI}(S;T)$, Bayesian risk $\mathrm{Br}(S;T)$ and complementary Bayesian risk $\overline{\mathrm{Br}}(S;T)$. In general, these indices are defined as follows for arbitrary random variables X and Y:

$$\begin{aligned}
H(X) &= -\sum_{x\in R(X)} \Pr(X=x)\log_2 \Pr(X=x);\\
I(X;Y) &= \sum_{(x,y)\in R(X)\times R(Y)} \left[\Pr(X=x,Y=y)\cdot \log_2 \frac{\Pr(X=x,Y=y)}{\Pr(X=x)\Pr(Y=y)}\right];\\
\mathrm{NMI}(X;Y) &= \frac{2\cdot I(X;Y)}{H(X)+H(Y)};\\
\mathrm{Br}(X;Y) &= 1-\sum_{x\in R(X)} \max_{y\in R(Y)} \Pr(X=x,Y=y);\\
\overline{\mathrm{Br}}(X;Y) &= 1-\mathrm{Br}(X;Y).
\end{aligned}$$

In the equations above, we assume that the range $R(X)$ and $R(Y)$ of random variables X and Y are finite, but however, these indices can be defined for more general settings: A random variable $X : \Omega \to R(X)$ is a measurable function from a probability space (Ω,Σ,p) to a measure space $(R(X),\mathcal{A},\mu)$, and a Radon-Nikodym derivative of $p\circ X^{-1}$ (probability density function) $f : R(X) \to \mathbb{R}$, if present, satisfies $\Pr[X \in A] \overset{\triangle}{=} p(X^{-1}(A)) = \int_A f d\mu$ for any $A \in \mathcal{A}$. Therefore, the information entropy $H(X)$ of X, for example, is defined by $H(X) = \int_{R(X)} -f\log_2 f d\mu$.

The Shannon information (or information content) of an event observing a value x for a random variable X is defined by $-\log_2 \Pr[X=x]$. It is interpreted as the quantity of information that the event carries, and the information entropy $H(X)$ is the mean across all the possible observables of X. Moreover, when we simultaneously observe

$X = x$ and $Y = y$, $-\log_2 \Pr[X = x] - \log_2 \Pr[Y = y] - (-\log_2 Pr[X = x, Y = y])$ quantifies the overlap of Shannon information between the events of $X = x$ and $Y = y$. The mutual information $I(X;Y)$ is the mean of the overlap, and therefore, quantifies overall correlational relation between observables of X and Y. In fact, we have:

- $I(X;Y) = 0$, if, and only if, X is independent of Y;
- $I(X;Y) = H(Y)$, if, and only if, Y is totally dependent on X, that is, $\Pr[Y = y \mid X = x]$ is either 0 or 1 for any x and y.

The normalized mutual information $\mathrm{NMI}(X;Y)$, on the other hand, is defined by the harmonic mean of $\frac{I(X;Y)}{H(X)}$ and $\frac{I(X;Y)}{H(Y)}$, and hence, takes values in $[0,1]$. We have:

- $\mathrm{NMI}(X;Y) = 0$, if, and only if, X is independent of Y;
- $\mathrm{NMI}(X;Y) = 1$, if, and only if, X and Y are isomorphic as random variables.

Bayesian risk $\mathrm{Br}(X;Y)$ also quantifies correlation of X to Y, which takes values in $\left[0, \frac{|R(Y)|-1}{|R(Y)|}\right]$. In contrast to mutual information, a smaller value of $\mathrm{Br}(X;Y)$ indicates a tighter correlation. This is why we use $\overline{\mathrm{Br}}(X;Y) = 1 - \mathrm{Br}(X;Y)$ in some cases. In particular, we have:

- $\mathrm{Br}(X;Y) = 0$, if, and only if, Y is totally dependent on X;
- $\overline{\mathrm{Br}}(X;Y) = 1$, if, and only if, Y is totally dependent on X.

The inequality below describes the relationship between $I(X;Y)$ and $\overline{\mathrm{Br}}(X;Y)$ [17]:

$$\begin{aligned} -\log_2 \overline{\mathrm{Br}}(X;Y) &\le H(Y) - I(X;Y) \\ &\le -\overline{\mathrm{Br}}(X;Y)\log_2 \overline{\mathrm{Br}}(X;Y) + \mathrm{Br}(X;Y)\log_2 \frac{\mathrm{Br}(X;Y)}{|R(Y)|-1}. \end{aligned} \tag{1}$$

In the remainder of this paper, we suppose that D is a dataset described by a feature set $\mathcal{F}$, which consists of only categorical features. Furthermore, unless otherwise noted, a feature f is supposed to be a member of $\mathcal{F}$, and a feature set S is a subset of $\mathcal{F}$.

3 Supervised Feature Selection Vs. Unsupervised Feature Selection

The most significant difference between feature selection in supervised learning and feature selection in unsupervised learning lies in whether class labels can be used as effective guides when selecting features. We will first review the literature on supervised feature selection.

The literature shows that the following four principles are commonly considered in designing supervised feature selection algorithms:

- Maintaining high class relevance;
- Reducing the number of selected features;
- Reducing the internal redundancy of selected features;
- Reducing the information entropy of selected features.

In the following illustration, we assume that S is a feature set selected by any feature selection algorithm from the entire feature set $\mathcal{F}$ that describes a dataset D. We also let C denote the random variable that yields class labels.

The class relevance of S represents the extent to which the features of S correlate to class labels and can typically be measured by the mutual information $I(S;C)$. In fact, $I(S;C)$ quantifies the part of the information content $H(C)$ of C that is also born by S. And hence, the class relevance $I(S;C)$ of S cannot exceed the entire information content $H(C)$ or the class relevance $I(\mathcal{F};C)$ of $\mathcal{F}$.

On the other hand, the purpose of feature selection is indeed to reduce the number of features to be used for explaining class labels. By its nature, the class relevance of selected features is a monotonically-increasing function with respect to the inclusion relation. In fact, for mutual information, we have $I(T;C) \leq I(S;C)$, if $T \subseteq S$. Therefore, the most fundamental problem of supervised feature detection can be stated as follows:

> The fundamental problem of supervised feature selection.
>
> Eliminate the maximum number of features while minimizing the resulting reduction of class relevance.

We have two important categories of features to eliminate or not to select.

Irrelevant features bear only a small amount of information content useful for explaining class labels. A feature f with small mutual information $I(f;C)$ is irrelevant.

Redundant features, on the other hand, bear content information that is mostly covered by the remaining features. For example, we suppose that S is a set of features selected tentatively. A redundant feature $f \in S$ makes $H(S) - H(S \setminus \{f\})$ sufficiently small. This implies that $I(S;C) - I(S \setminus \{f\};C)$ is also sufficiently small.

In the literature, the well-known feature selection algorithm MRMR (Minimum Redundancy and Maximum Relevance) [11] tries to eliminate irrelevant features and redundant features. To determine a feature f to add to the tentative solution S, it intends to evaluate the index of

$$b(f,S) = (I(S,f;C) - I(S;C)) - \frac{I(S;f)}{|S|},$$

which quantifies a balance between contribution to class relevance and increase of redundancy by adding f to S. Computing $b(f,S)$ is, however, costly, and MRMR uses the following approximation.

$$b(f,S) \approx b'(f,S) = I(f;C) - \frac{\sum_{f' \in S} I(f;f')}{|S|}. \tag{2}$$

Algorithm 1 describes MRMR. The asymptotic time complexity of MRMR is estimated by $O(k^2|\mathcal{F}||D|)$.

MRMR is one of the most well-known feature selection algorithms and in fact has been not only intensively studied but also used widely in practice [2,5,10,13,14,20,23,23,26]. CFS [6] is another feature selection algorithm that is widely used in practice.

Algorithm 1. Mrmr [11].

```
Require: A dataset $D$ described by $\mathcal{F} \cup \{C\}$; and $k < |\mathcal{F}|$.
Ensure: A feature subset $S \subseteq \mathcal{F}$ with $|S| = k$.
Let $S = \emptyset$.
while $|S| < k$ do
    Let $f \in \arg\max\{b'(f, S) \mid f \in \mathcal{F} \setminus S\}$.
    Let $S = S \cup \{f\}$.
end while
Return $S$.
```

It is also based on the same principle as Mrmr but uses a different formula than Eq. (2) to evaluate a balance of class relevance and interior redundancy.

These algorithms, however, encounter the problem of ignoring feature interaction [24]. We say that two or more features *mutually interact* when each individual feature has only low class relevance, but the group of these features has high class relevance. The aforementioned algorithms, which only evaluate the information entropy of individual features, cannot detect mutual feature interaction, and are likely to discard interacting features, which can result in a loss of class relevance.

Zhao et al. [24] pointed out the importance of this issue and proposed Interact, the first algorithm that evaluates feature interaction and realizes practical time-efficiency at the same time. Interact has led to the development of many algorithms including Lcc [15,16], which improve Interact in both accuracy (when used with classifiers) and time-efficiency. Interact and Lcc use the complementary Bayesian risk $\overline{\mathrm{Br}}(S;C)$ to measure class relevance of S. Equation (1) describes the correlational relation between $\overline{\mathrm{Br}}(S;C)$ and $I(S;C)$. Lcc takes a single hyper-parameter t, which specifies a lower limit of class relevance of the output feature set. Algorithm 2 describes the algorithm of Lcc. Also, Cwc [15,18] is equivalent to LCC with $t = 1$.

Algorithm 2. Lcc [15,16].

```
Require: A dataset $D$ described by a feature set $\mathcal{F} \cup \{C\}$; and a threshold $t \in [0,1]$.
Ensure: A minimal feature subset $S \subseteq \mathcal{F}$ with $\overline{\mathrm{Br}}(S;C) \geq t\overline{\mathrm{Br}}(\mathcal{F};C)$.
1: Number the features of $\mathcal{F}$ so that $f_1, \ldots, f_{|\mathcal{F}|}$ are in a decreasing order of $\mathrm{NMI}(f_i;C)$.
2: Let $S = \mathcal{F}$ and $i = n_{|\mathcal{F}|}$.
3: while $i \geq 1$ do
4:     Let $j = \arg\min\{j \mid j \in [1, i+1], \overline{\mathrm{Br}}(S \setminus \mathcal{F}[j,i];C) \geq t\overline{\mathrm{Br}}(\mathcal{F};C)\}$.
5:     Let $S = S \setminus \mathcal{F}[j,i]$ and $i = j - 2$.
6: end while
7: Return $S$.
```

Unlike Mrmr, Interact and Lcc evaluate class relevance of S by $\overline{\mathrm{Br}}(S;C)$ without using approximation based on evaluation of individual features. By this, they not only can eliminate irrelevant and redundant features but also can incorporate feature interaction into feature selection. Although computing $\overline{\mathrm{Br}}(S;C)$ is more costly than using the approximation, Lcc drastically improves time efficiency by taking advantage of binary search when searching a feature f_j to select in Step 4 of Algorithm 2. Due to this, the asymptotic time complexity of Lcc is $O(|\mathcal{F}||D|\log|\mathcal{F}|)$, and its practical time efficiency is significantly high.

The principle of reducing information entropy is loosely related to the principle of reducing number of features, although they are not equivalent to each other. Explanation of phenomena using fewer features is more understandable for humans, while explanation using features with smaller entropy is more efficient from a information-theoretical point of view.

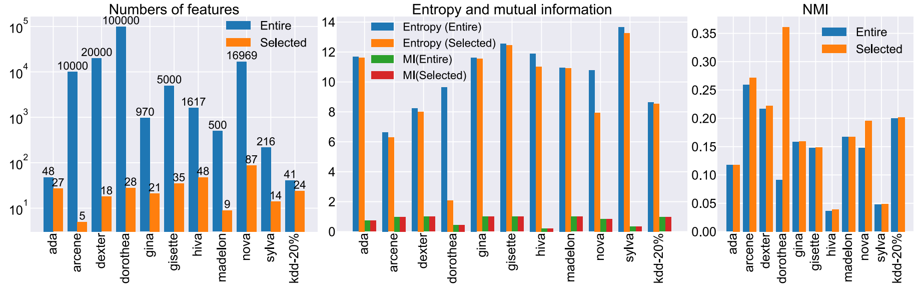


Fig. 1. Before and after of supervised feature selection.

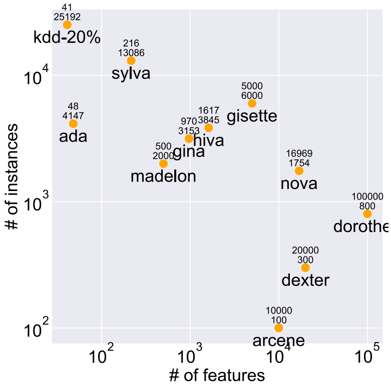


Fig. 2. Eleven datasets used in our experiment.

Figure 1 shows the results of feature selection by CWC when applied to the 11 datasets described in Fig. 2 and below. These datasets are relatively large and encompass a wide range in terms of the numbers of features and instances. They are taken from the literature: five from NIPS 2003 Feature Selection Challenge, five from WCCI 2006 Performance Prediction Challenge, and one from KDD-Cup Challenge. Since our interest in this paper lies in the selection of categorical features, the continuous features included in the datasets are categorized into five equally-long intervals before feature selection. The instances of all of the datasets are annotated with binary labels.

From the left chart of Fig. 1, we first notice that CWC has selected a reasonably small number of features for all the datasets. For example, while the dataset DOROTHEA originally includes 100,000 features, CWC has selected only 28 features.

The middle chart of Fig. 1 shows that the dataset except DOROTHEA and NOVA include many redundant features and only a few irrelevant features. In fact, we see only small losses of the information entropy between the entire features $\mathcal{F}$ (blue) and the selected features S (orange). In contrast, CWC eliminates many irrelevant features from DOROTHEA and NOVA: The gaps between the information entropy of $\mathcal{F}$ and S are large for these datasets. By the nature of CWC, there is no loss in mutual information. It is interesting to note that the normalized mutual information scores of S are better than $\mathcal{F}$ for some of the datasets (the right chart of Fig. 1). In particular, the extents of improvement for DOROTHEA and NOVA are significant. Since the normalized mutual information quantifies the extent of the identity between two random variables, the selected features S can explain class labels more confidently than the entire features $\mathcal{F}$. This emphasizes the importance of feature selection in addition to the advantages in model construction and the improved efficiency of post-feature-selection learning.

So far, we have shown the four most important guiding principles in designing supervised feature selection algorithms. Among them, the principle of maintaining high class relevance cannot be used in unsupervised feature selection, because we do not have class labels in unsupervised learning. The issue here lies in the fact that there is a trade-off between maintaining high class relevance and each of the last three principles, and supervised feature selection selects features that realize an appropriate balance between class relevance and the other indices. Without class relevance, the other three indices do not mutually constrain, and it turns out that selecting no features is always the optimal answer. Thus, the most important question for realizing successful unsupervised feature selection is how to find one or more principles that constrain the other three principles without leveraging class labels. We propose an answer for this question in Sect. 5.

4 Feature Value Selection Vs. Feature Selection

The advantages of feature value selection over feature selection are two-fold:

1. Feature value selection is likely to explain the same phenomena using factors with less information content. This means that the explanation is more efficient and more accurate.
2. We sometimes use the term *modeling* to indicate selecting of a small number of effective explanatory variables from a larger pool of possible variables to explain objective variables. Using feature values as explanatory variables improves the concreteness of the explanation.

4.1 More Efficient Explanation of Phenomena

For formalization, we first introduce a binarized feature and a binarized feature set. The purpose is to reduce feature value selection to feature selection by converting a feature value into a binary vector using one-hot encoding.

Definition 1. *For a value $v \in R(f)$, $v_{@f}$ denotes a binary feature such that for an instance $x \in D$, $v_{@f}(x) = 1$ if $f(x) = v$; otherwise, $v_{@f}(x) = 0$.*

Definition 2. *For a set S of features, we determine $S^b = \{v_{@f} \mid f \in S, v \in R(f)\}$.*

As explained in Sect. 2, a binarized feature $v_{@f} \in S^b$ is canonically viewed as a random variable defined over the sample space $\prod_{v_{@f} \in S^b} \mathbb{Z}_2$, where D determines an empirical probability measure. $\mathbb{Z}_2$ denotes $\{0, 1\}$. In particular, we can convert a dataset D into a new dataset D^b, which consists of the same instances but is described by $\mathcal{F}^b$. Thus, we can equate feature value selection on a dataset D to feature selection on D^b.

The entire features $\mathcal{F}$ of the dataset of the next example can explain class labels necessarily and sufficiently. We see that we can also select feature values $S \subset \mathcal{F}^b$ that completely explain the class labels as well. Although the cardinality of S is the same as that of $\mathcal{F}$, $H(S)$ is significantly smaller than $H(\mathcal{F})$.

Example 1. We consider n features $f_1, \ldots, f_n$, whose values are arbitrary b-dimensional binary vectors (b-bit-long natural numbers), that is, $R(f_i) = \mathbb{Z}_2{}^b$. For an instance $x \in D$, we determine its class label $C(x)$ by

$$C(x) = \left(\sum_{i=1}^{n} 2^{i-1} \cdot 0_{@f_i}(x) \right) \bmod L,$$

where L is an odd number that determines the number of classes. To illustrate, we further assume that f_i is independent of any other f_j, and the associated probability distribution is uniform. Because the class labels of x rely on all of $f_1(x), \ldots, f_n(x)$, the answer of feature selection for this dataset is unique and must be $\{f_1, \ldots, f_n\}$. For the same reason, the answer of feature value selection must be $\{0_{@f_1}, \ldots, 0_{@f_n}\}$. While it is evident that

$$I(f_1, \ldots, f_n; C) = I(0_{@f_1}, \ldots, 0_{@f_n}; C) = H(C)$$

holds, the information entropy $H(f_1, \ldots, f_n) = nb$ is significantly greater than

$$H(0_{@f_1}, \ldots, 0_{@f_n}) = n \left(b2^{-b} - (1 - 2^{-b}) \log_2 (1 - 2^{-b}) \right) \approx n(b+1)2^{-b},$$

when b is not small. This means that, $0_{@f_1}, \ldots, 0_{@f_n}$ can explain the class C with the same accuracy but significantly more efficiently than $f_1, \ldots, f_n$.

The following general mathematical results justify the result of Example 1.

Theorem 1. *For disjoint feature sets S and T in $\mathcal{F}$, $H(S, T) = H(S^b, T)$ holds.*

Proof. For $\boldsymbol{v} \in R(S) = \prod_{f \in S} R(f)$, we determine $\boldsymbol{v}^b \in R(S^b) = \prod_{v@f \in S^b} \mathbb{Z}_2$ by $v_{@f}(\boldsymbol{v}^b) = 1 \Leftrightarrow f(\boldsymbol{v}) = v$. When we let $\mathcal{D} = \{\boldsymbol{v}^b \mid \boldsymbol{v} \in R(S)\} \subset R(S^b)$, we have the following for arbitrary $\boldsymbol{w} \in R(S^b)$ and $\boldsymbol{u} \in R(T)$:

$$\Pr[S^b = \boldsymbol{w}, T = \boldsymbol{u}] = \begin{cases} \Pr[S = \boldsymbol{v}, T = \boldsymbol{u}], & \text{if } \boldsymbol{w} \in \mathcal{D}, \text{ that is,} \exists \left(\boldsymbol{v} \in \prod_{f \in S} R(f)\right) [\boldsymbol{w} = \boldsymbol{v}^b]; \\ 0, & \text{if } \boldsymbol{w} \notin \mathcal{D}. \end{cases}$$

Hence, the assertion follows:

$$\begin{aligned} H(S,T) &= \sum_{\boldsymbol{v} \in R(S)} \sum_{\boldsymbol{u} \in R(T)} -\Pr[S = \boldsymbol{v}, T = \boldsymbol{u}] \log_2 \Pr[S = \boldsymbol{v}, T = \boldsymbol{u}] \\ &= \sum_{\boldsymbol{w} \in \mathcal{D}} \sum_{\boldsymbol{u} \in R(T)} -\Pr[S^b = \boldsymbol{w}, T = \boldsymbol{u}] \log_2 \Pr[S^b = \boldsymbol{w}, T = \boldsymbol{u}] \\ &\quad + \sum_{\boldsymbol{w} \in R(S^b) \setminus \mathcal{D}} \sum_{\boldsymbol{u} \in R(T)} -\Pr[S^b = \boldsymbol{w}, T = \boldsymbol{u}] \log_2 \Pr[S^b = \boldsymbol{w}, T = \boldsymbol{u}] \\ &= H(S^b, T). \end{aligned}$$

□

The following corollaries to Theorem 1 explain Example 1.

Corollary 1. *For $S \subseteq \mathcal{F}$, $H(S) = H(S^b)$ holds.*

Corollary 2. *For feature subsets S and T in $\mathcal{F}$, $I(S;T) = I(S^b;T)$ holds.*

Proof. Theorem 1 implies

$$I(S;T) = H(S) + H(T) - H(S,T) = H(S^b) + H(T) - H(S^b,T) = I(S^b;T). \square$$

By the monotonicity properties of information entropy and mutual information, if $S' \subset S^b$, we have $H(S') \leq H(S)$ and $I(S';T) \leq I(S;T)$. Example 1 is the case where $H(S') \ll H(S)$ holds, while $I(S';T) = I(S;T)$ holds, for $S = \{f_1, \ldots, f_n\}, S' = \{0_{@f_1}, \ldots, 0_{@f_n}\}$ and $T = \{C\}$.

4.2 More Concrete Modeling

Feature value selection explains how features contribute to the determination of class labels more clearly. Even if a feature f is selected through feature selection, not all of the possible values of f necessarily contribute to the determination equally. In particular, only a small portion of values may be useful for explaining class labels.

For example, an Intrusion Protection System (IPS) tries to detect a small portion of packets generated for malicious purposes out of the large volume of packets that are transmitted in networks. Based on the information of the detected malicious packets, IPS tries to take effective measures to protect a system. To a packet, multiple headers of protocols such as TCP, IP and IEEE 802.x are attached. The information born by these headers is the main source of information for IPS. For example, a TCP header includes a *Destination Port* field, and a value of this field usually specifies what application will receive this packet and will execute particular functions as a result of the reception. Since malicious attackers target particular vulnerable applications, knowing what potion numbers are correlated to malicious attacks will allow an IPS to take more accurate countermeasures than only knowing that values of the destination port field.

5 Fast Unsupervised Feature Selection

The basis of our proposed algorithm was presented in [19]. We now explain our improvements to this algorithm using explainability.

To review, [19] provides a useful basis for developing an efficient unsupervised feature value selection algorithm:

- It leverages the principle that every instance must be explained by at least one selected feature value. This principle constrains the minimization of the three remaining factors (feature value count, internal redundancy, and information entropy, as discussed in Sect. 3) and guarantees that at least one meaningful solution exists.
- It incorporates the algorithmic framework of Lcc [16, 18] by leveraging binary search, which gives it significantly high time efficiency;
- It contains one hyperparameter for excluding feature values below a threshold of information entropy;

We build on this algorithm by adding two features:

1. We introduce the concept of *explainability* as a substitute for the concept of class relevance that plays a central role in supervised feature selection.
2. We add two hyperparameters: the minimum of collective explainability across all selected feature values, and a minimum explainability for each individual feature value.

5.1 Explainability-Based Unsupervised Feature Selection

Supervised learning provides an effective guide for feature selection in the form of class relevance scores. There are several measures of class relevance: Mrmr [26] and Cfs [6] use mutual information $I(S;C)$ following [3], while Interact [24] and Lcc [16] deploy the complementary Bayesian risk $\overline{\mathrm{Br}}(S;C)$.

On the other hand, unsupervised feature selection has no class labels to measure the relevance of features to class labels. As a substitute, then, we introduce *explainability*. In [19], the support of a set of feature values is defined as follows:

Definition 3. ([19]). *For $S \subseteq \mathcal{F}^b$, the support of S is defined by*

$$\mathrm{supp}_D(S) = \{x \in D \mid \exists(v_{@f} \in S)[f(x) = v]\}.$$

The support $\mathrm{supp}_D(S)$ consists of the instances that possess at least one feature value included in S, or, in other words, are explained by the feature values in S.

Definition 4. *The* explainability *of S is determined by*

$$\mathfrak{X}_D(S) = \frac{|\mathrm{supp}_D(S)|}{|D|}.$$

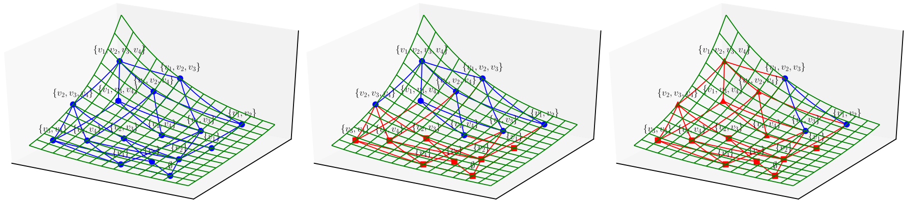
Having defined explainability, we can formally define ξ-*explainability-based unsupervised feature value selection* as follows.

ξ-Explainability-Based Unsupervised Feature Value Selection (ξ-EUFVS).

Given an unlabeled dataset D described by a feature set $\mathcal{F}$ and the lower limit ξ of explainability, find $S \subseteq \mathcal{F}^b$ that minimizes $H(S)$ or $|S|$, or both if possible, subject to the condition of $\mathfrak{X}_D(S) \geq \xi$.

As explained in Sect. 3, although the information entropy $H(S)$ and the size $|S|$ are loosely correlated, minimizing one does not necessarily mean minimizing the other. Also, $H(S)$ is important from an explanation efficiency point of view, while $|S|$ affects the understandability of the obtained model by humans. Thus, the aforementioned formalization leaves some ambiguity in terms of objective functions, but however, this does not significantly matter in practice, because finding exact solutions to the problem of ξ-EUFVS is likely to be computationally impossible. When solving it approximately, the aforementioned loose correlation between $H(S)$ and $|S|$ helps us reach a reasonable balance between them.

We see how explainability performs as a substitute for class relevance using Fig. 3. To illustrate, we assume that $\mathcal{F}^b$ consists of only four values v_1, v_2, v_3, v_4.



(a) The Hasse diagram of $\{v_1, v_2, v_3, v_4\}$

(b) Restriction by ξ-explainability

(c) Restriction by individual explainability

Fig. 3. Search space of EUFVS.

The chart (a) depicts the Hasse diagram of $\mathcal{F}^b$, which is a directed graph (V_H, E_H) such that V_H is the power set of $\mathcal{F}^b$, and $(S,T) \in V_H \times V_H$ is in E_H, if, and only if, $S \supset T$ and $|S| - |T| = 1$ hold. The height of a plot of $S \subseteq \mathcal{F}^b$ represents the magnitude of $H(S)$. We will start at the top node of $\mathcal{F}^b$ and will search the node that minimizes $H(S)$ and/or $|S|$ by following directed edges downward. If the search space is the entire Hasse diagram, it is evident that we can stop when we reach the bottom node that represents the empty set $\emptyset$. This solution is indeed trivial and meaningless. Thus, we need an appropriate restriction on the search space. For supervised feature selection, the principle of maintaining high class relevance narrows down the search space, because small feature sets can have only low class relevance (for example, $I(\emptyset; C) = 0$ holds), and such nodes are eliminated from the search space. As a result, we can reach a non-trivial meaningful node in the search space. The condition that the explainability $\mathfrak{X}_D(S)$ is no smaller than the predetermined threshold ξ has the same effect. Like mutual information, the explainability index is monotonous with respect to the inclusion relation of feature value sets: if $T \subseteq S \subseteq \mathcal{F}^b$, $\mathfrak{X}_D(T) \leq \mathfrak{X}_D(S)$ holds. This means that, if S is out of the search space, that is, $\mathfrak{X}_D(S) < \xi$ holds, any $T \subseteq S$ is out of the search space as well. In particular, $\text{supp}_D(\emptyset) = \emptyset$ and $\mathfrak{X}_D(\emptyset) = 0$ hold. Chart (b) of Fig. 3 depicts this.

The sets $T \subseteq \mathcal{F}^b$ with $\mathfrak{X}_D(T) < \xi$ are displayed in red, and we see that there is more than one minimal selection S in the sense that $\mathfrak{X}_D(S) \geq \xi$ holds but $\mathfrak{X}_D(T) < \xi$ holds for arbitrary $T \subsetneq S$. All of these minimal nodes S comprise the set of candidate solutions to the ξ-EUFVS problem.

As the threshold ξ for the entire explainability $\mathfrak{X}_D(S)$ increases, the resulting search space becomes narrower, and the border has a higher altitude. In other words, the threshold ξ moves the border of the search space in the vertical direction.

In addition to the threshold ξ, we introduce a different threshold t for individual explainability $\mathfrak{X}_D(v)$ for individual feature value $v \in \mathcal{F}^b$. This threshold t constrains the search space so that a node in the space includes only feature values v whose individual explainability $\mathfrak{X}_D(v)$ is not smaller than t. In contrast to the threshold ξ to collective explainability, this threshold has the effect of moving the border of the search space in the horizontal direction. For example, in Fig. 3(c), we assume that $\mathfrak{X}_D(v_1) > \mathfrak{X}_D(v_2) > \mathfrak{X}_D(v_3) \geq t > \mathfrak{X}_D(v_4)$. Then, the subgraph displayed in blue is the search space determined in combination with ξ.

The introduction of this threshold t can be justified as follows.

- For example, if $\mathcal{F}$ includes a feature f, which yields a unique identifier for each instance of D, the support of any feature value $v_{@f}$ is a singleton, and hence, its individual explainability is positive but minimum. Evidently, selecting a unique identifier $v_{@f}$ is of no help for understanding the dataset. Although this example is extreme, in general, a feature value whose support is a very small set of instances lacks generality, and it is not desirable to include it in selection.
- As already explained, the threshold t for individual explainability moves the border of the search space in the horizontal direction, while the threshold ξ for entire explainability does in the vertical direction. By combining these two thresholds, we can move the border of the search space in both the vertical and horizontal directions, and hence, we will have multiplicative flexibility to define the range of solutions to the EUFVS problem.

At last of this subsection, we note the relation between $\mathfrak{X}_D(v)$ and $H(v)$. Since v is a feature value, and therefore, is binary as a random variable, we have

$$H(v) = -\mathfrak{X}_D(v)\log_2 \mathfrak{X}_D(v) - (1 - \mathfrak{X}_D(v))\log_2(1 - \mathfrak{X}_D(v)).$$

Since the function $F(x) = -x\log_2 x - (1-x)\log_2(1-x)$ is an increasing function for $x \in [0, \frac{1}{2}]$, if $\mathfrak{X}_D(v) \leq \frac{1}{2}$, the threshold t on $\mathfrak{X}_D(v)$ is equivalent to the threshold $F(t)$ on $H(v)$. The algorithm presented in [19] takes a threshold to $H(v)$ as a hyperparameter. When we assume that $\mathfrak{X}_D(v) \leq \frac{1}{2}$, these two definitions of hyperparameters are equivalent to each other.

5.2 The Algorithm

Algorithm 3 describes the algorithm that we propose in this paper. Due to the monotonicity property of $\mathfrak{X}_D(S) \leq \mathfrak{X}_D(T)$ for $S \subseteq T$, we can take advantage of a binary search to find the next feature value to leave in S (Step 5). As a result, the algorithm is significantly fast as shown in Sect. 6.1.

Algorithm 3. Explainability-based Unsupervised Feature Value Selection (EUFVS).

```
Require: An unlabeled dataset $D$ described by $\mathcal{F}$; a threshold $\xi \in \left[\frac{1}{2}, 1\right]$; a threshold $t \in \left[0, \frac{1}{2}\right]$.
Ensure: A minimal feature value set $S \subseteq \mathcal{F}^b$.
Let $\overline{S} = \mathcal{F}^b \setminus \{v_{@f} \in \mathcal{F}^b \mid \mathfrak{X}_D(v_{@f}) \le t\}$.
Number the feature values of $\overline{S}$ so that $\overline{S} = \{v_1, \ldots, v_{|\overline{S}|}\}$ and $\mathfrak{X}_D(v_i) \ge \mathfrak{X}_D(v_j)$ for $i < j$.
Let $l = 0$ and $S = \overline{S}$.
while $l < |\overline{S}|$ do
    Let $k = \max\{j \mid \mathfrak{X}_D(S \setminus \overline{S}[l+1, j]) \ge \xi, j = l, \ldots, |\overline{S}|\}$ by binary search.
    Let $S = S \setminus \overline{S}[l+1, k]$ and $l = k+1$.
end while
return $S$.
```

The time complexity of Algorithm 3 can be estimated as follows: the complexity of computing $\mathfrak{X}_D(v_i)$ and $\mathfrak{X}_D(\mathcal{F}^b[i, |\mathcal{F}^b|])$ for all i is $O(|\mathcal{F}^b| \cdot |D|)$; By updating $\mathfrak{X}_D(S \cap \mathcal{F}^b[1, l])$ whenever we update l, $\mathfrak{X}_D(S \setminus \mathcal{F}^b[l+1, j]) \ge \xi$ can be investigated in $O(|D|)$-time, and the average complexity to execute the while loop is estimated by $O((\log_2 |\mathcal{F}^b|)^2 \cdot |D|)$.

6 Empirical Performance Evaluation

We conducted three experiments with EUFVS. The first assessed the basic performance of EUFVS, while the second and third applied EUFVS to real-world data, specifically tweets and electricity consumption.

6.1 Basic Performance Evaluation

To measure EUFVS's performance compared to other algorithms, we tried it on the 11 datasets from well-known machine learning challenges shown in Fig. 2. We used well-known datasets to ensure our results were comparable to other algorithms tested on these datasets. Our goal was to discover how accurately and quickly EUFVS could build feature value sets that explained these datasets' labels, with the labels removed.

In the experiment, we set the threshold on collective explainability to $\xi = 1$ and changed the threshold on individual explainability t on each iteration so that the maximum value would not exceed 5% of the total number of instances in each datsets.

Runtime Performance. Figure 4 describes the runtime of Algorithm 3 in milliseconds for three typical datasets: KDD-20% with significantly many instances, DOROTHEA with significantly many features, and GISETTE with both many instances and many features (Fig. 2). The scores include only the search time. The runtime was under 100 milliseconds for all datasets, except for when we used very small thresholds. The longest run was GISETTE with a threshold of $t = 0$, which took only 2,500 ms.

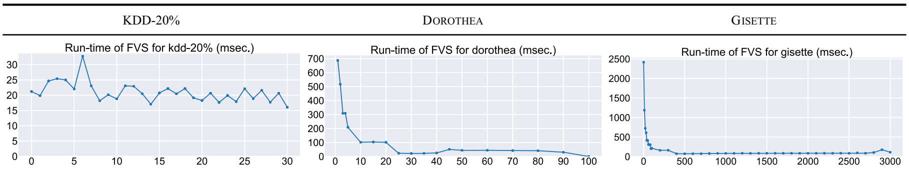


Fig. 4. Runtime in milliseconds for different t values (x-axis).

Selection Performance. Several affinities appear in the results of nine of these eleven datasets. We will describe these using the examples of GISETTE and SYLVA (The left and middle columns of Fig. 5).

1. All 11 datasets consisted of labelled data, which provided a ground truth to test against. Our goal was to see how well EUFVS could produce feature sets that explained the labels without the labels for guidance, so we removed the labels from the datasets.
   Even so, it found feature value sets that explain the labels well. In fact, $I(S;C)$ remains close to $I(\mathcal{F};C)$, until t exceeds a certain limit. This property is significant evidence that our algorithm has an excellent ability to select *appropriate feature values*, because the dataset labels are a *perfect summary* of the datasets.
2. When t exceeds the said limit, $I(S;C)$ rapidly decreases. In other words, the different selected feature value sets represent different views of the datasets.
3. As t increases, $I(S;C)$ and $H(S)$ synchronously decrease. This implies that our algorithm eliminates non-redundant and relevant feature values after it has eliminated all the redundant feature values.
4. $H(S)$ remains very close to its upper bound of $H(\mathcal{F})$ (the orange line) until t reaches the said limit. By contrast, the number of feature values selected decreases rapidly immediately when t increases. This may imply that an overwhelming majority of feature values v with small $H(v)$ are redundant.
5. The number of the selected feature values approaches the number of the features selected by CWC (the green line). This implies that approximately one value for each feature selected by CWC is truly relevant to class labels.

The evaluation result of KDD-20% are also interesting. KDD-20% is a dataset of network packet headers gathered by intrusion detection software. Each instance is labelled as either "normal" or "anomalous". Unlike in the other datasets, the score of $H(S)$ moves around half of $H(\mathcal{F})$, while $I(S;C)$ remains close to $I(\mathcal{F};C)$. In fact, KDD-20% and ADA are the only datasets that could exhibit higher NMI$(S;C)$ than NMI$(\mathcal{F};C)$. With high $I(S;C)$ and low $H(S)$, the feature values selected could have good classification capability when used with a classifier. Also, it is surprising that the number of feature values selected is smaller than 30, when they show the highest score of NMI$(S;C)$. The figure is significantly lower than the 225 feature values that CWC selects for this dataset, and hence, could provide a much more interpretable model.

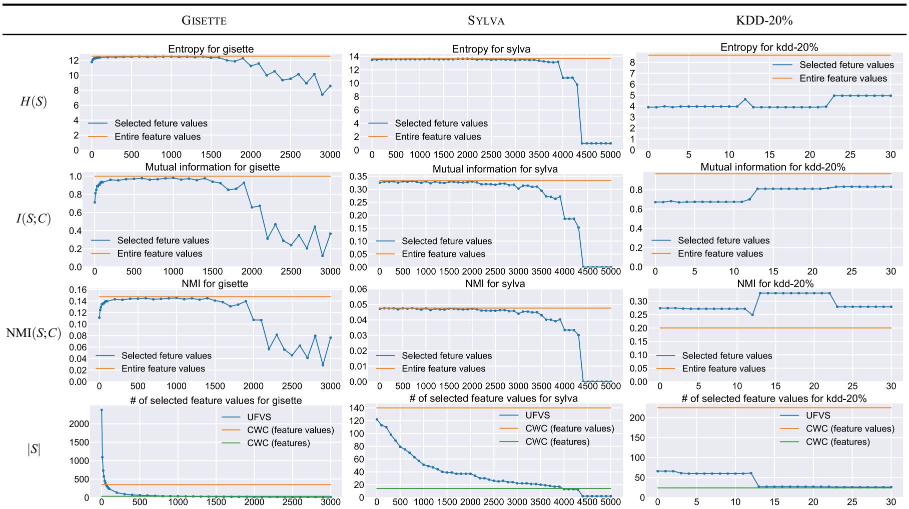


Fig. 5. Comparison in $H(S)$, $I(S;C)$, $\mathrm{NMI}(C;S)$ and $|S|$ for typical three datasets.

Classification Performance. The experimental results on the (normalized) mutual information scores for the selection results of EUFVS imply that the selection results can accurately predict the dataset's class labels even though the inputs into EUFVS do not include them.

To investigate their potential classification performance, we ran experiments with the three typical classifiers: CART, Naïve Bayes (NB) and Supprt Vector Machine (C-SVM) classifiers. When we use the SVM classifier, we project instances to points in higher-dimensional spaces using the RBF kernel.

We evaluate classification performance using averaged accuracy scores obtained through five-fold cross-validation. Optimal hyperparameter values are determined using grid search based on five-fold cross-validation scores on training data, executed at each fold execution to compute a single accuracy score. For comparison, we also run the same experiments on feature value sets selected by MRMR [11].

Figure 6 describes the results for the three typical datasets of GISETTE, SYLVA and KDD-20%. From the charts, we can observe the following.

- For four deatasets including GISETTE, the accuracy scores for EUFVS encompass a relatively wide range, and some of them are compatible with the results of MRMR.
- For six datasets including SYLVA, the accuracy scores for the selection results of EUFVS varies within a relatively narrow range, and are sometimes better and sometimes worse than those for MRMR. Overall, their classification performance appear to be compatible with the selection results of MRMR.
- The results for KDD-20% is surprising, since the accuracy scores significantly better than MRMR. To be precise, for all classifiers and all feature value sets selected by MRMR, the accuracy scores fall within the range between 0.5 and 0.6. This will make us conclude that the selection results of MRMR are not useful for the purpose of classification. In contrast, the accuracy scores for EUFVS distributes in a narrow range around the value of 0.9.

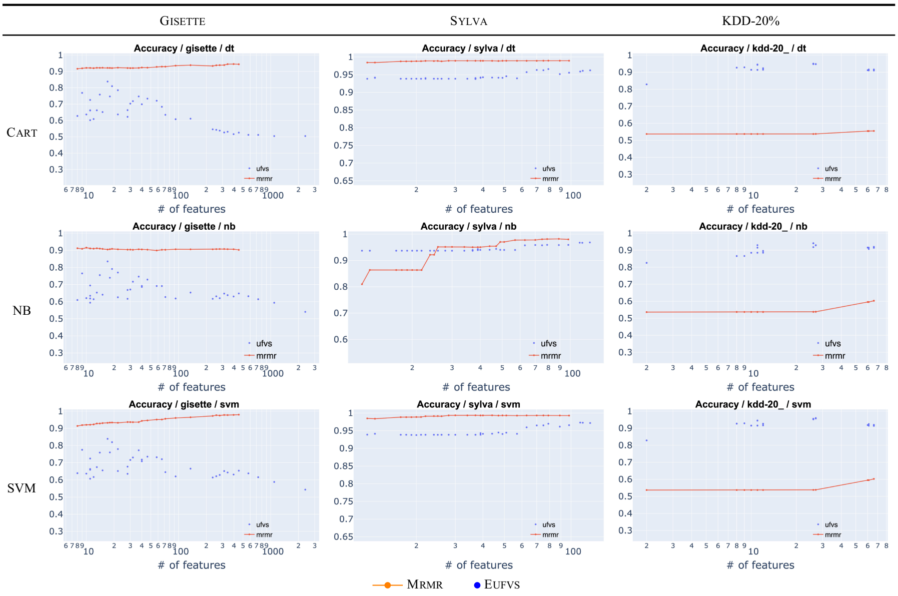


Fig. 6. Comparison in accuracy by Cart, NB and SVM classifiers for typical three datasets.

Phase Transition by Change of Threshold. For the experiments using the 11 labeled datasets, we investigate the differences between the feature value sets selected by EUFVS for different threshold values t. For this purpose, we leverage the distance derived from the Jaccard coefficient and the k-means algorithm, a distance-based clustering algorithm.

The Jaccard index between two sets S and T is defined by $J(S,T) = \frac{|S \cap T|}{|S \cup T|}$. It is known that $J(S,T)$ is positive definite, and hence, $d_J(S,T) = \sqrt{2 - 2J(S,T)}$ is identical to the Euclid distance in some Euclidean space (reproducing kernel Hilbert space) between the projections of S and T in the space. This is derived from the well-known cosine formula. In particular, when a finite number of sets $S_1, \ldots, S_n$ are given, they can be projected into a common n-dimensional Euclidean space, and their coordinates in the space is computed as follows. We let $J = [J(S_i, S_j)]_{i,j}$ be the Gram matrix. Since the Jaccard index is positive definite, the Schur decomposition of J is as follows with $\lambda_i \geq 0$:

$$J = U^{\top} \mathrm{diag}(\lambda_1, \ldots, \lambda_n) U.$$

When we let $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n] = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) U$, $\boldsymbol{v}_i$ gives a coordinate of S_i in an n-dimensional space.

Here, we let $S_1, \ldots, S_n$ be the feature value sets selected by EUFVS for n different thresholds. Because we can concretely project them into an n-dimensional space, we can apply the k-means clustering algorithms to the projections in plural times for different cluster count k.

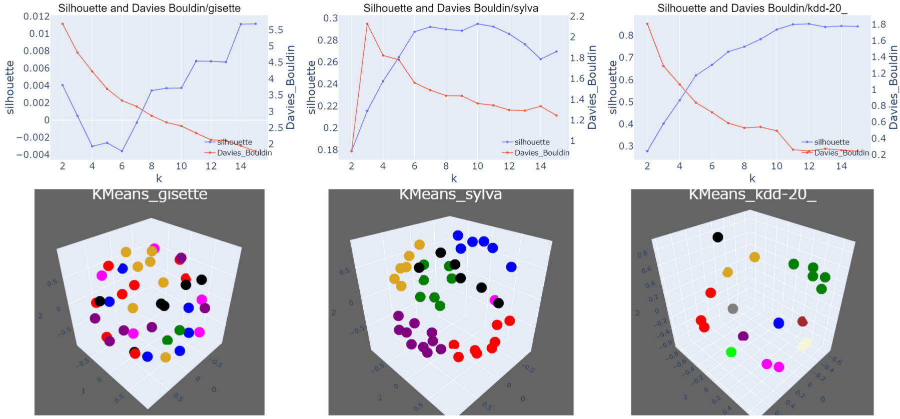
The upper row of Fig. 7 shows the transition of the silhouette coefficient and the Davies Bouldin index as k increases, and based on it, we determine the optimal k for each dataset. Note that a greater silhouette coefficient and a smaller Davies Bouldin index indicate better clustering. At the same time, We also prefer to use the smallest possible k. With these constraints, we determine $k = 15$ for GISETTE, $k = 7$ for SYLVA and $k = 11$ for KDD-20% as optimal cluster counts.

With these values of k determined individually for the three datasets, the charts in the lower row of Fig. 7 depict the distributions of plots of the feature value sets in a three-dimensional space after reducing the dimensionality by MDS. For SYLVA and KDD-20%, we see that clustering has performed well, and clusters correspond exactly to consecutive intervals of thresholds. This implies that changing threshold parameter values results in a continuous change of viewpoint over these datasets. For GISETTE, no clear correspondence between clusters and thresholds was found.

6.2 Experiments on Analysis of Twitter Data

This experiment shows an example of applying EUFVS to the analysis of real Twitter data.

GISETTE | SYLVA | KDD-20%



GISETTE: No clear correspondence found.

SYLVA:

Color	Thresholds	Color	Thresholds
Purple	0 – 900	Red	1000 – 1700
Black	1800 – 2300	Goldenrod	2400 – 2900
Green	3000 – 3600	Blue	3700 – 4300
Magenda	4400 – 5000		

KDD-20%:

Color	Thresholds	Color	Thresholds
Blue	0 – 12	Green	13 – 22
Magenda	23 – 150	Red	200 – 800
Black	1000 – 1350	Goldenrod	1400 – 2000

Fig. 7. Clustering of selected feature sets based on the jaccard index.

The data used in this experiment include 24,142 tweets posted on March 1, 2020 from 9:00 PM to 9:30 PM. This dataset is a set of tweets sent by users during the time who were tweeting about the COVID-19 pandemic. All of the users tweeted at least once about COVID-19, but not all of the tweets in the data set were about COVID-19. We extracted keywords from each tweets using the MeCab morphological analyzer, resulting in a matrix of 24,142 tweets (instances) × 49,342 unique words (features) as a feature table.

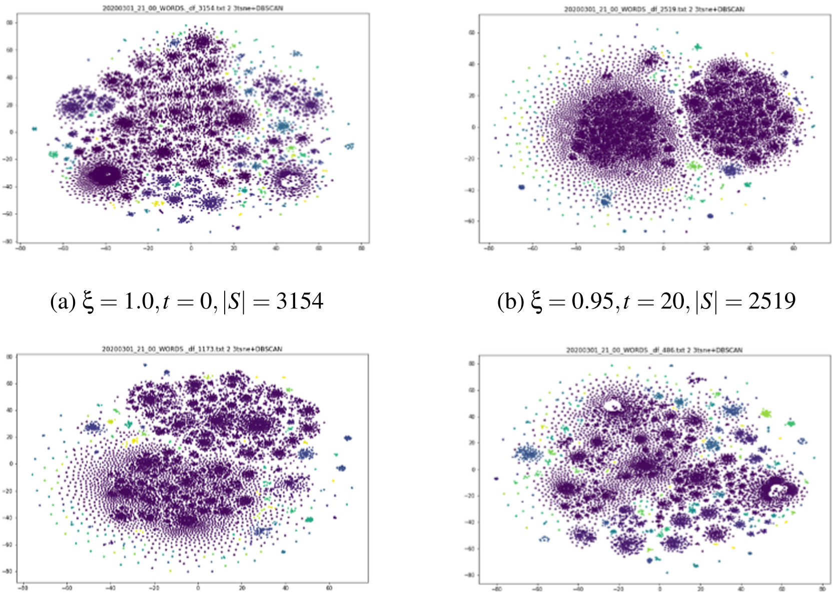
We performed feature selection using EUFVS multiple times with different parameter settings. Then, the UFVS results were compressed into two dimensions by TSNE (Manhattan distance) and were clustered by DBSCAN. Figure 8 shows the clustering results with different parameter settings: (a) $\xi = 1.0, t = 0$; (b) $\xi = 0.95, t = 20$; (c) $\xi = 0.9, t = 40$; and (d) $\xi = 0.95, t = 80$.

Figure 8(a) shows the clustering result with $\xi = 1.0$ and $t = 0$. After EUFVS, 3154 features remained. Several relatively large clusters can be seen. We have observed some clusters that represent COVID-19-related topics.

Figure 8(b) shows the clustering result with $\xi = 0.95$ and $t = 20$. After EUFVS, 2519 features remained. The set of Tweets was clearly divided into right and left sides. They may represent the characteristics of some Tweets groups. In fact, coronavirus-related clusters can be observed on the right side.

Figure 8(c) (1173 features, $\xi = 0.90$, $t = 40$) is also divided into two major groups, showing a trend similar to that of Fig. 8(b). On the other hand, Fig. 8(d) (486 features, $\xi = 0.95$, $t = 80$) shows a pattern similar to Fig. 8(a). In addition, the ring-shaped Tweet set can be seen in Figs. 8(a) and (d). The significance of these clusters will be a subject of future research.

Changing the parameter settings of EUFVS allows us to view the set of tweets from different perspectives. Our future work will explore these applications.



(a) $\xi = 1.0, t = 0, |S| = 3154$

(b) $\xi = 0.95, t = 20, |S| = 2519$

(c) $\xi = 0.9, t = 40, |S| = 1173$

(d) $\xi = 0.95, t = 80, |S| = 486$

Fig. 8. Clustering of tweets.

6.3 Experiments on Analysis of Electricity Consumption Data

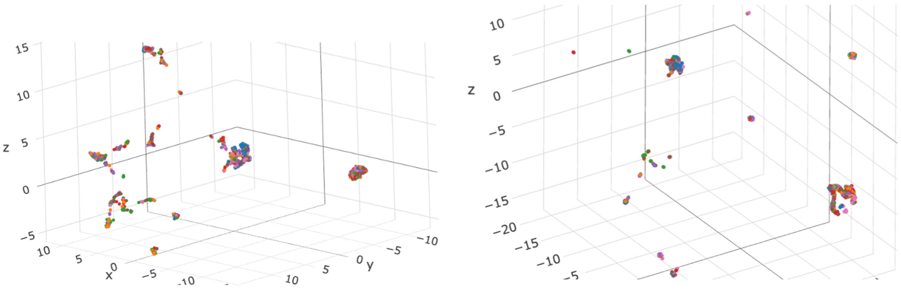
In this experiment, we apply EUFVS to analyzing power consumption in university classrooms. The on/off binary data of electricity consumption in each classroom is recorded in 30-min. time slots over one semester of 15 weeks. By concatenating daily slices (9 am to 6 pm) of this data, we obtain a table where one instance is one classroom on one day, and each feature is a one-hour time slot from 9 am to 6 pm. The data size is 3782 instances and 19 features.

We performed a number of feature selection trials using EUFVS with varying parameter settings of ξ and t. Here, we show two examples. Figures 9 includes 3D visualizations of the data after feature selection by EUFVS. For a dimensionality reduction algorithm, we used UMAP with the Euclidean distance. Figure 9(a) shows the results of EUFVS with the parameter settings of $\xi = 1.0$ and $t = 1660$. Thirteen features remained in this parameter setting. There are two dense clusters on the right and center of the figure, and a spreading, string-like cluster on the left side of the figure. The present data include classrooms in two buildings. Classrooms in the same building tend to cluster together, which reflects different power consumption patterns in different buildings.

Figure 9(b) shows the results of EUFVS with $\xi = 0.95$ and $t = 2000$. Five features remain in this parameter setting. Two dense clusters similar to those in Fig. 9(a) can be found, but the spreading, string-like cluster disappears. On the other hand, the structures of the two dense clusters of Fig. 9(b) are easily visible.

Figure 10 shows an enlarged view of the two dense clusters. These clusters also have a string-like structure. Furthermore, the structure is found to be sequentially linked from the first week to the 15th week. The instances are plotted in different colors for each day of the week, and the same color appears periodically, which is a manifestation of that.

As described above, by changing the parameter settings, we can obtain the visualizations suitable for the spreading, string-like cluster and suitable for the dense clusters. This is due to the design of EUFVS, which allows us to check a large number of various solutions through lightweight calculations.



(a) $\xi = 1.0, t = 1660$

(b) $\xi = 0.95, t = 2000$

Fig. 9. UMAP 3D view.

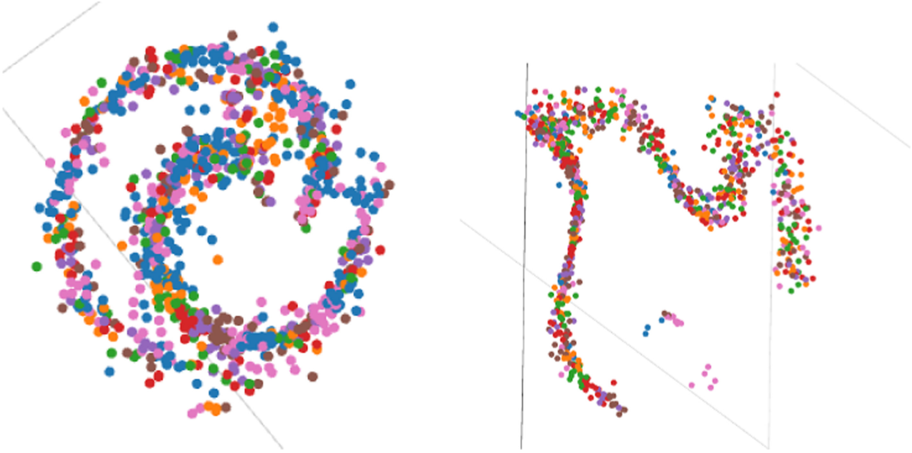


Fig. 10. Enlarged views of two dense clusters in Fig.9.

7 Conclusion

The intractability of unsupervised feature selection is caused by the indefiniteness problem: any dataset can contain multiple theoretically-correct solutions, but there is no metric for picking the best one. Instead of attempting to find a metric, EUFVS accepts the indefiniteness problem and works around it by speeding up the feature selection process. Because EUFVS is a fast, tunable algorithm, it empowers a human being to select the best result from many options.

Acknowledgements. This work was partially supported by the Grant-in-Aid for Scientific Research (JSPS KAKENHI Grant Numbers 16K12491 and 17H00762) from the Japan Society for the Promotion of Science.

References

1. Almuallim, H., Dietterich, T.G.: Learning boolean concepts in the presence of many irrelevant features. Artif. Intell. **69**(1–2), 279–305 (1994)
2. Angulo, A.P., Shin, K.: mRMR+ and CFS+ feature selection algorithms for high-dimensional data. Appl. Intell. **49**(5), 1954–1967 (2019). https://doi.org/10.1007/s10489-018-1381-1. https://doi.org/10.1007/s10489-018-1381-1
3. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Netw. **5**(4), 537–550 (1994)

4. Cai, D., Zhang, C., He, X.: Unsupervised feature selection for multi-cluster data. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010), pp. 333–342 (2010)
5. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. In: Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003, pp. 523–528 (2003)
6. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: ICML 2000 (2000)
7. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: Advances in Neural Information Processing Systems (NIPS 2005), pp. 507–514 (2005)
8. Li, Z., Liu, J., Yang, Y., Zhou, X., Liu, H.: Clustering-guided sparse structural learning for unsupervised feature selection. IEEE Trans. Knowl. Data Eng. **26**(9), 2138–2150 (2014)
9. Liu, H., Shao, M., Fu, Y.: Consensus guided unsupervised feature selection. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2016), pp. 1874–1880 (2016)
10. Mohamed, N.S., Zainudin, S., Othman, Z.A.: Metaheuristic approach for an enhanced MRMR filter method for classification using drug response microarray data. Expert Syst. Appl. **90**, 224–231 (2017). https://doi.org/10.1016/j.eswa.2017.08.026. http://www.sciencedirect.com/science/article/pii/S0957417417305638
11. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
12. Qian, M., Zhai, C.: Robust unsupervised feature selection. In: Proceedings of 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013), pp. 1621–1627 (2013)
13. Radovic, M., Ghalwash, M., Filipovic, N., Obradovic, Z.: Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinform. **18**(1), 9 (2017). https://doi.org/10.1186/s12859-016-1423-9
14. Senawi, A., Wei, H., Billings, S.A.: A new maximum relevance-minimum multicollinearity (mrmmc) method for feature selection and ranking. Pattern Recognit. **67**, 47–61 (2017). https://doi.org/10.1016/j.patcog.2017.01.026
15. Shin, K., Fernandes, D., Miyazaki, S.: Consistency measures for feature selection: a formal definition, relative sensitivity comparison, and a fast algorithm. In: 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011). pp. 1491–1497 (2011)
16. Shin, K., Kuboyama, T., Hashimoto, T., Shepard, D.: sCWC/sLCC: highly scalable feature selection algorithms. Information **8**(4), 159 (2017)
17. Shin, K., Xu, X.: Consistency-based feature selection. In: 13th International Conference on Knowledge-Based and Intelligent Information & Engineering System (2009)
18. Shin, K., Kuboyama, T., Hashimoto, T., Shepard, D.: Super-CWC and super-LCC: super fast feature selection algorithms. Big Data **2015**, 61–67 (2015)
19. Shin, K., Okumoto, K., Shepard, D., Kuboyama, T., Hashimoto, T., Ohshima, H.: A fast algorithm for unsupervised feature value selection. In: 12th International Conference on Agents and Artificial Intelligence (ICAART 2020), pp. 203–213 (2020). https://doi.org/10.5220/0008981702030213
20. Vinh, L.T., Thang, N.D., Lee, Y.K.: An improved maximum relevance and minimum redundancy feature selection algorithm based on normalized mutual information. In: 2010 10th IEEE/IPSJ International Symposium on Applications and the Internet, July 2010. https://doi.org/10.1109/saint.2010.50. http://dx.doi.org/10.1109/SAINT.2010.50
21. Wei, X., Cao, B., Yu, P.S.: Unsupervised feature selection on networks: a generative view. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2016), pp. 2215–2221 (2016)

22. Wei, X., Cao, B., Yu, P.S.: Multi-view unsupervised feature selection by cross-diffused matrix alignment. In: Proceedings of 2017 International Joint Conference on Neural Networks (IJCNN 2017), pp. 494–501 (2017)
23. Zhang, Y., Ding, C., Li, T.: Gene selection algorithm by combining reliefF and mRMR. BCM Genomics **9**(2), 1–10 (2008)
24. Zhao, Z., Liu, H.: Searching for interacting features. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI 2007), pp. 1156–1161 (2007)
25. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: Proceedings of the 24th International Conference on Machine Learning (ICML 2007), pp. 1151–1157 (2007)
26. Zhao, Z., Anand, R., Wang, M.: Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform, August 2019