IVFS: Simple and Efficient Feature Selection for High Dimensional Topology Preservation

Xiaoyun Li, Chengxi Wu, and Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th ST. Bellevue WA, 98004, USA
{lixiaoyun996, wuchenxi2013, pingli98}@gmail.com

Abstract

Feature selection is an important tool to deal with high dimensional data. In unsupervised case, many popular algorithms aim at maintaining the structure of the original data. In this paper, we propose a simple and effective feature selection algorithm to enhance sample similarity preservation through a new perspective, topology preservation, which is represented by persistent diagrams from the context of computational topology. This method is designed upon a unified feature selection framework called IVFS, which is inspired by random subset method. The scheme is flexible and can handle cases where the problem is analytically intractable. The proposed algorithm is able to well preserve the pairwise distances, as well as topological patterns, of the full data. We demonstrate that our algorithm can provide satisfactory performance under a sharp sub-sampling rate, which supports efficient implementation of our proposed method to large scale datasets. Extensive experiments validate the effectiveness of the proposed feature selection scheme.

Introduction

High dimensional data becomes more and more common in machine learning applications, e.g., computer vision, natural language processing and gene selection. In many cases, the curse of dimensionality leads to costly computation and a less comprehensible model. Therefore, feature selection is one of the standard data preprocessing methods. The most widely used approach is filter method, where each feature is individually assigned a score according to some statistical measure, and those with highest scores are selected. Supervised filter methods include t-test (Guyon and Elisseeff 2003), mutual information (Peng, Long, and Ding 2005), correlation (Yu and Liu 2003), etc. Meanwhile, there are many unsupervised filter algorithms based on similarity and manifold preservation. LaplacianScore algorithm (He, Cai, and Niyogi 2006) are proposed to choose features according to a nearest neighbor graph. Features that are smoothest on the graph are regarded as most capable to represent the local manifold structure of full data. Spectral Feature Selection (SPEC) (Zhao and Liu 2007) aims at separating the samples into clusters using the spectrum of pairwise similarity graph. Multocluster Feature Selection (MCFS) (Cai, Zhang, and He 2010) considers combining spectral embedding and $l_1$-regularized regression by assuming a multi-clustered structure of the data.

Although filter methods are fast to implement, the performance can sometimes be unsatisfactory. In recent years, many learning-based embedded feature selection methods are proposed. The idea of most embedded approaches is to optimize a selection (or weight) matrix with some sparsity regularization. The objective function can be designed to achieve similarity preservation. For instance, Similarity Preserving Feature Selection (SPFS) (Zhao et al. 2013) learns a linear transformation $W$ by minimizing

$$
\min_W \|XW(XW)^T - S\|_F^2 + \lambda \|W\|_{2,1}, \tag{1}
$$

where $X$ is the data, $S$ is the sample similarity matrix and $W$ is weight matrix. Features with largest row $l_2$ norm of $W$ are selected. SPFS mainly considers preserving similarity globally. An improved version of SPFS is called Global and Local Structure Preservation Feature Selection (GLSPFS) (Liu et al. 2014), whose objective modifies (1) as

$$
\min_W \|XW(XW)^T - S\|_F^2 + \mu \cdot \text{tr}(W^T LXW) + \lambda \|W\|_{2,1},
$$

where $L$ is a locality representation graph. This way, local structure is also considered. Other unsupervised embedded algorithms include NDFS (Li et al. 2012), RUFS (Qian and Zhai 2013), FSASL (Du and Shen 2015), SOGFS (Nie, Zhu, and Li 2016), AHLFS (Zhu et al. 2017), UPFS (Li et al. 2018b), DGUFS (Guo and Zhu 2018), etc.

Typically, existing similarity preserving feature selection schemes face two major challenges:

- **High dimensionality.** As will be shown in our experiments, many methods actually perform not as good in high dimensions in terms of distance preserving, possibly due to lack of consideration of feature interactions, or huge number of parameters to optimize. This may result in non-robust or unstable solutions.
- **Large sample size.** In most of the embedded methods introduced above (GLSPFS, SOGFS, etc.), the algorithm involves (repeatedly) matrix decomposition and/or inverse, which runs very slowly with large sample size.
Topology preserving feature selection

Many embedded methods are associated with clustering or nearest neighbor graph embedding, which in some sense address more on the local manifold structure. In this paper, we look at similarity preservation problem from a new view: topology. In recent years, topological data analysis (TDA) has been shown as a powerful tool for various learning tasks (Hofer et al. 2017). The merit of TDA comes from its capability to encode all the topological information of a point set $X$ by persistent diagram, denoted as $D(X)$ herein. In Euclidean space, $D(X)$ is tightly related to pairwise sample distances. This motivates us to consider similarity preservation from a TDA perspective. If we hope to select covariates $X_F$ to preserve the topological information of the original data, the persistent diagram $D(X_F)$ generated by $X_F$ should be close to the original diagram $D(X)$. More precisely, TDA offers a new space (of persistent diagrams) in which we compare and preserve the distances. We refer this property as “topology preservation”, which is very important especially when applying TDA after feature selection. In existing literature, however, topology preservation has not been considered yet.

Our contributions

We develop a simple but powerful filter based feature selection method that achieves topology preservation:

- We propose a unified scheme called Inclusion Value Feature Selection (IVFS) that extends the idea of random subset evaluation to unsupervised case. It generalizes to several well-known supervised methods, i.e., random forest (Breiman 2001) and random KNN (Li, Harner, and Adjeroh 2011; Räsänen and Pohjalainen 2013).

- We use IVFS to achieve topology (persistent diagram) preserving feature selection, based on theories from TDA. By reinforcing the scalability, this algorithm not only perform very well in high dimensions, but also run efficiently with large sample size.

- We conduct extensive experiments on high dimensional datasets to justify the effectiveness of IVFS. The results suggest that IVFS is capable of preserving the exact topological signatures (as well as the pairwise distances), making it a suitable tool for TDA and many other applications.

IVFS: A Unified Feature Selection Scheme

First, we provide a unified framework based on random subset evaluation, which is flexible, universally applicable, and generalizes to several existing algorithms. In this section, we provide a detailed discussion on this general scheme.

Problem formulation and notations

The data matrix is denoted by $X \in \mathbb{R}^{n \times d}$ with $n$ samples and $d$ covariates.$^1$ Our goal is to find the best $d_0 < d$ features according to some criteria, associated with a loss function $\mathcal{L}$. Denote $\mathcal{I} = \{1, 2, ..., d\}$ the indices of all features, and $\mathcal{M}_d = \{\sigma \in \mathcal{I}^d : \sigma_i \neq \sigma_j \text{ for } \forall i \neq j\}$ the set of all size-$d$ subsets of $\mathcal{I}$. The goal is to minimize the objective function

$$\min_{F \in \mathcal{M}_{d_0}} \mathcal{L}_F \triangleq \min_{F \in \mathcal{M}_{d_0}} \mathcal{L}(X_F). \quad (2)$$

Our algorithm relies on repeatedly sampling random subset $F$ of arbitrarily $d$ features (not necessarily equal to $d_0$), equipped with a subset score function $s(F, X) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ which assigns score to each selected feature by evaluating the chosen random subset. In principle, a high score should correspond to a small loss.

IVFS: a filter method for feature selection

The individual feature score, which serves as the filter of the unified selection scheme, essentially depends on the subset score function $s(\cdot)$ defined above.

Definition 1. Suppose $1 \leq \tilde{d} \leq d$. The Inclusion Value of feature $f \in \mathcal{I}$ at dimension $\tilde{d}$ associated with $s(\cdot)$ is

$$IV_{\tilde{d}}(f) = \frac{\sum_{\sigma \in \mathcal{M}^f_{\tilde{d}}} s_\sigma(f)}{(\tilde{d}-1)},$$

where $\mathcal{M}^f_{\tilde{d}} = \{\sigma \in \mathcal{M}_{\tilde{d}} : f \in \sigma\}$ is the collection of subsets with size $\tilde{d}$ that contains feature $f$, and $s_\sigma(f)$ is the score assigned to feature $f$ by computing $s(\sigma;X)$.

Intuitively, the inclusion value illustrates how much gain in score a feature $f$ could provide on average, when it is included in the feature subset of size $\tilde{d}$. Our feature selection scheme is constructed based on inclusion value estimation, as summarized in Algorithm 1. We call it Inclusion Value Feature Selection (IVFS). Roughly speaking, the algorithm selects features with highest estimated inclusion value, which is derived via $k$ random sub-samplings of both features and observations. One benefit is that IVFS considers complicated feature interactions by evaluating subset of features together in each iteration.

Special cases. The IVFS scheme includes several popular methods as special cases based on different score function (i.e., the inclusion value).

- Permutation importance. For each feature in a random subset $F$, if we set $s_F(f)$ as the difference between the performance (e.g., classification accuracy, regression mean squared error) using $F$ and the performance when feature vector $f$ is randomly permuted, then IVFS becomes the feature selection algorithm via supervised random forest permutation importance (Strobl et al. 2008).

- RKNN. When we set $s_F(f) = -\mathcal{L}_F$, $\forall f \in F$, where $\mathcal{L}$ is the KNN classification error rate, IVFS becomes supervised RKNN (Li, Harner, and Adjeroh 2011). Note that for random forest, the score function $s_F(f)$ is different for each feature $f \in F$, while in RKNN, all the features in a random subset share a same score.
Algorithm 1: IVFS scheme for feature selection

Input: Data matrix $X \in \mathbb{R}^{n \times d}$; Number of subsets $k$; Number of features used for each subset $d$; Number of samples for each subset $n$; Target dimension $d_0$

Initialize: Counters for each feature $c_f = 0$, $i = 1, \ldots, d$; Cumulative score for each feature $S_f = 0$

for $t = 1$ to $k$

Randomly sample a size $\tilde{d}$ feature set $\mathcal{F} \in \mathcal{M}_{\tilde{d}}$

Randomly sub-sample $X^{\text{sub}} \in \mathbb{R}^{n \times \tilde{d}}$, with $\tilde{n}$ observations and features in $\mathcal{F}$

for $f$ in $\mathcal{F}$

Update counter $c_f = c_f + 1$

Update score $S_f = S_f + s(f)$

Set $S_i = \frac{S_i}{c_i}$ for $i = 1, 2, \ldots, d$

Output: Select top $d_0$ features with highest score

Analysis of IVFS

Theorem 1 (Asymptotic $k$). Suppose $k \to +\infty$, $\tilde{n} = n$, $s_\sigma(f)$ has finite variance $\forall f \in \mathcal{I}$, and the $IV_\tilde{d}$ for different features are all distinct, then IVFS algorithm will select top $d_0$ features with highest $IV_{\tilde{d}}$ with probability $1$.

Proof. The algorithm is equivalent to finding the top $d_0$ features with the largest

$$\hat{IV}_{\tilde{d}}(f) = \frac{\sum_{\sigma \in \hat{C}_d} s_\sigma(f)}{|\hat{C}_d|},$$

where $\hat{C}_d = \{\sigma \in \hat{C}_d : f \in \sigma\}$, and $\hat{C}_d$ is the collection of all $k$ chosen random subsets. By central limit theorem, when $k \to \infty$, we have for some $\tau$ and $\forall f$,

$$\frac{\hat{IV}_{\tilde{d}}(f) - IV_{\tilde{d}}(f)}{\sqrt{k}} \to N(0, \tau^2).$$

Let $\delta$ be the difference between the $d_0$-highest $IV_{\tilde{d}}$ and the $(d_0 + 1)$ highest, then for $\forall f$, for any $\epsilon > 0$ there exists a $K$ such that when $k > K$, the probability of $|\hat{IV}_d(f) - IV_{\tilde{d}}(f)| > \delta/2$ is less than $\epsilon$. Taking $\epsilon \to 0$ and $k \to \infty$, the theorem is proved.

In the following, we look at the case where $s_\mathcal{F}(f)$ are equal for every $f \in \mathcal{F}$ (e.g., RKNN). As a result, we may re-write the score function as $s(f)$ evaluated on subsets. The next assumption on monotonicity appears commonly in feature selection literature, similar in spirit to (Narendra and Fukunaga 1977; Foroutan and Sklansky 1987), etc.

Assumption 1 (Monotonicity). There exists a $d$-dimensional set $\Omega \in \mathcal{M}_d$ such that $\forall \mathcal{F}, \mathcal{F'} \in \mathcal{M}_d$, if $(\mathcal{F} \cap \Omega) \subseteq (\mathcal{F'} \cap \Omega)$, then $s(\mathcal{F}) \leq s(\mathcal{F'})$.

Basically, this assumption says that there is a subset $\Omega$ of “dominant features”: For any two subsets $\mathcal{F}, \mathcal{F'}$ with same size, if $(\mathcal{F} \cap \Omega)$ is contained in $(\mathcal{F'} \cap \Omega)$, then the score of $\mathcal{F'}$ is no smaller than that of $\mathcal{F}$. It turns out that this dominant set is indeed optimal, and it is also the solution that IVFS converges to in large $k$ limit.

Theorem 2 (Optimality). Under Assumption 1, we have

$$\Omega = \arg \max_{\mathcal{F} \in \mathcal{M}_d} s(\mathcal{F}),$$

and $\Omega$ is the set of $d$ features with the highest $IV_{\tilde{d}}$.

Proof. It suffices to show that for $\forall f \in \Omega$, $g \notin \Omega$, $IV_{\tilde{d}}(f) \geq IV_{\tilde{d}}(g)$. We have by assumption

$$IV_{\tilde{d}}(f) - IV_{\tilde{d}}(g) = \frac{1}{d} \left( \sum_{f \in \mathcal{F} \in \mathcal{M}_{\tilde{d}-1}} s(\mathcal{F} \cup \{f\}) - \sum_{g \notin \mathcal{F}} s(\mathcal{F} \cup \{g\}) \right) + \frac{1}{d} \sum_{f \notin \mathcal{F} \in \mathcal{M}_{\tilde{d}-1}} s(\mathcal{F} \cup \{f\}) - s(\mathcal{F} \cup \{g\}) \geq 0,$

which proves the second argument. Since for $\forall \mathcal{F} \in \mathcal{M}_d$, $(\mathcal{F} \cap \Omega) \subseteq (\Omega \cap \Omega)$, we have $\Omega = \arg \max_{\mathcal{F} \in \mathcal{M}_d} s(\mathcal{F})$.

Together with Theorem 1, we know that if we set $\tilde{d} = d_0$, under Assumption 1, IVFS would converge to the minimal score feature set. For instance, in the case of RKNN, the selected features would minimize the KNN error rate.

Choosing $\tilde{d}$. In practice (when $k \ll \binom{n}{d_0}$), we are actually drawing random samples from the population, and use feature scores as estimation of true inclusion values. An interesting fact is that, IVFS actually uses $IV_{\tilde{d}}$ to estimate $IV_{d_0}$. This makes sense since 1) we expect features with high $IV_{\tilde{d}}$ to have high $IV_{d_0}$ as well, and 2) we care more about the rank of feature scores rather than the exact values. Hence, setting $d > d_0$ may not defeat the model performance.

Choosing $\tilde{n}$. We also have a parameter $\tilde{n}$ which controls the number of random observations for each subset. A relatively small $\tilde{n}$ is extremely helpful to accelerate the algorithm for scalable implementation on large datasets. In existing methods, however, sub-sampling is not commonly used.

- In the original proposal of random forest (Breiman 2001), the authors suggested to set $\tilde{n} = n$ with replacement. Later on, a popular variant (Geurts, Ernst, and Wehenkel 2006) chose to disable sub-sampling procedure. In general, in all variants of random forest, sub-sampling is not recommended (Tang, Garreau, and von Luxburg 2018).

- In RKNN (Li, Harner, and Adjeroh 2011), the author did not consider sub-sampling training points, either.

For supervised learning, the above phenomenon seems reasonable, since sub-sampling the training data may harm the learning capacity of each random subset, especially with high dimensions. However, as shall be seen from next sections, when applying IVFS for the purpose of unsupervised topology preservation, we can choose a very small $\tilde{n}$ without loosing much capacity. This makes IVFS a strong candidate for dealing with large scale datasets. In general, a good choice of $\tilde{n}$ depends on the specific problem.
Advantages of IVFS. On a high level, IVFS has the following nice features:

- Intuitive formulation, no complicated computation.
- IVFS can well handle the problems where the optimization problem is very hard to solve (or intractable), as long as the loss function can be computed efficiently.
- IVFS can be applied to large datasets efficiently by using a small $\tilde{n}$, which is feasible for some applications.

In the following sections, we design a topology preserving feature selection algorithm by combining IVFS framework and ideas from topological data analysis (TDA).

Preliminaries on Computational Topology

In this section, we provide some intuition to several important concepts in computational topology. Interested readers are referred to (Edelsbrunner and Harer 2010) for more detailed introduction. A $p$-dimensional simplex is defined as

$$\gamma_p = \{\theta_0 x_0 + \ldots + \theta_p x_p | \theta_i > 0 \forall i, \sum_{i=0}^{p} \theta_i = 1\},$$

where $x_0, \ldots, x_p$ are affinely independent points in $\mathbb{R}^p$. For instance, a 1-simplex is a line segment, and a 2-simplex is a triangle, etc. A simplicial complex $\mathcal{C}$ is then formed by gluing simplices in different dimensions together. In Euclidean space, the most commonly used complex is the Vietoris-Rips complex, with an example given in Figure 1. It is formed by connecting points with distance smaller than a given threshold $\alpha$. If we gradually increase $\alpha$ from 0 to $\infty$, the number of edges will increase from 0 to $n^2$ eventually. The distance associated with each edge, is called the filtration for Rips complex. As $\alpha$ increases, topological features with different dimension (e.g., 0 for connected components, 1 for loops, 2 for voids, etc.) will appear and disappear. We call the pair of birth and death time (the $\alpha$ value) of a $p$-dimensional topological feature as a $p$-dimensional persistent barcode. The $p$-dimensional persistent diagram is a multiset of all these barcodes. An example persistent diagram is plotted in Figure 1 right panel. Note that we can always normalize the filtration function to be bounded in $[0,1]$.

Figure 1: Left panel: an example of Rips complex. Points that are close are connected. Dots, lines and triangles forms 0, 1 and 2-dimensional complex respectively. Right panel: an example of persistent diagram. Each red point represents a (birth, death) time of a topological feature. The green line is a threshold: only barcodes that are stable (points above the green line) are included in the final barcode set.

Distances between persistent diagrams

Recall that our objective involves minimizing the difference between persistent diagrams. The following two distances measures between diagrams are widely used in TDA.

Definition 2. For two persistent diagrams $\Sigma$ and $\Gamma$, define Wasserstein distance ($w_\infty$) and Bottleneck distance ($w_\infty$)

$$w_\infty(\Sigma,\Gamma) = \inf_{\phi} \left( \sum_{x \in \Sigma} \|x - \phi(x)\|_{\infty} \right)^\frac{1}{p},$$

$$w_\infty(\Sigma,\Gamma) = \inf \sup \{ \|\phi(x)\|_{\infty} | x \in \Sigma \},$$

where $\phi$ is taken over all bijections $\Sigma \rightarrow \Gamma$ and $p, q \in \mathbb{N}$.

Objective function

Now we are ready to formally state our objective function. Recall the notations $\mathcal{I} = \{1, 2, \ldots, d\}$, and $\mathcal{M}\boldsymbol{_{\mathcal{I}}} = \{\sigma \in \mathcal{I}^d | \sigma_i \neq \sigma_j \text{ for } \forall i \neq j\}$. To achieve topology preserving feature selection, we minimize following loss function,

$$\min_{\mathcal{F} \in \mathcal{M}\boldsymbol{_{\mathcal{I}}}} w_*(\mathcal{D}(X), \mathcal{D}(X_{\mathcal{F}})), \quad (3)$$

where $w_*$ denotes Wasserstein or bottleneck distance. This way, we find the subset that best preserves the topological signatures of original data. However, the mapping between feature space and persistent diagram is so sophisticated that analytical approach to (3) is hard to derive. This is exactly the circumstance where IVFS is particularly effective.

Note that the computational cost to generate persistent diagram is non-negligible even for data of moderate size. Hence, directly applying IVFS with objective loss (3) would
be extremely slow. To this end, we leverage from the stability property of persistent diagrams to propose an alternative solution. We re-state the theorems as below.

**Theorem 3.** (Chazal, De Silva, and Oudot 2014) Suppose $X$ is the point set, $f$ and $f'$ are two Lipschitz functions. Let $D(X, f)$ and $D(X, f')$ denote the persistent diagram built upon $X$ using filtration function $f$ and $f'$, respectively. Under some technical conditions, the bottleneck and Wasserstein distances are bounded by

$$w_{\infty}(D(X, f), D(X, f')) \leq C_1 \|f - f'\|_{\infty},$$

$$w_p^p(D(X, f), D(X, f')) \leq C_2 \|f - f'\|_{C^2}^p,$$

where $C_1$, $C_2$ and $C_3$ are a universal constant independent of $X$, and $\| \cdot \|_{\infty}$ refers to the infinite norm.

In words, the theorem says that when we change filtration from $f$ to $f'$, the change in persistent diagrams would be bounded linearly in the $\|f - f'\|_{\infty}$ for Bottleneck distance, and polynomially for Wasserstein distance. Since the filtration for Rips complex is the pairwise distances, we can distance, and polynomially for Wasserstein distance. Since the

**Methods and tuning parameters**

We compare several popular similarity preserving methods:

- **SPEC**: We use the second type algorithm which performs the best, according to (Zhou and Liu 2007).
- **MCFS**: As guided in (Cai et al., 2010), we run MCFS with number of clusters $M = \{5, 10, 20, 30\}$.
- **GLSPFS**: It is a embedded method that learns to approximate the sample similarity matrix and will serve as a major baseline. Following (Liu et al. 2014), we chose the parameter combinations of $\mu = \{0, 10^{-1}, 3\}$, $\lambda = \{10^{-1}, 3\}$.

**IVFS-$l_{\infty}$ and variants**: We try following combinations: $d = \{0.1 : 0.1 : 0.5\} \times d$, $n = \{100, 0.1n, 0.3n, 0.5n\}$. We run experiments with $k = 1000, 3000, 5000$. All the reported results are averaged over 5 repetitions.

**Evaluation Metrics**

We compare each method by various widely adopted metrics that can well evaluate the quality of selected feature set.

**KNN accuracy.** Following (Zhou and Liu 2007; Liu et al. 2014), etc., we test local structure preservation by KNN classification. Each dataset is randomly split into 80% training sets and 20% test set, on which the test accuracy is computed. We repeat the procedure 10 times and take the average accuracy. For number of neighbors, we adopt $K \in \{1, 3, 5, 10\}$ and report the highest mean accuracy.

**Distances between persistent diagrams.** For $X$ and $X_f$, we compute the 1-dimensional persistent diagram $D$ and $D_F$ with $\alpha = 0.5$ and drop all barcodes with existing time less than $\epsilon = 0.1$. Wasserstein ($w_1^1$) and Bottleneck ($w_{\infty}$) distances are computed between the diagrams.

**Norms between distance matrix.** When the purpose of feature selection is to preserve the sample distance (or similarity), one straightforward measure should be the change between distance matrices $D$ and $D_F$. We compute $\|D - D_F\|_1$, $\|D - D_F\|_2$ and $\|D - D_F\|_{\infty}$ to evaluate the closeness of these two matrices.

**Running time.** We compare the running time for a single run, with fixed parameter setting: MCFS: $k = 10$, GLSPFS: $\mu = 1$, $\lambda = 1$, and IVFS: $k = 1000$, $n = 0.1n$, $d = 0.3d$.

**Results**

**Overall performance.** Table 1 summarizes the results. All the datasets are high-dimensional, and Isolet, RELATHE, COIL20 have relatively larger size with around 1500 samples. For algorithms with tuning parameters, we report the best result among all parameters and number of selected features for each metric. From Table 1, we observe:

- **IVFS-$l_{\infty}$** provides smallest $w_1^1$, and $w_{\infty}$ on almost all datasets. Moreover, the $l_1$, $l_2$ and $l_{\infty}$ norms are significantly reduced on all the data—The distance and topology preserving capability is essentially improved.

- On all datasets, IVFS-$l_{\infty}$ also beats other methods in terms of KNN accuracy, which indicates its superiority on supervised tasks and local manifold preservation.

We also tested normalized mutual information (NMI) in the experiment. The pattern was very similar to KNN accuracy.
Robustness. We plot $\|D - D_\hat{F}\|_2$ and $w_\infty(D, D_\hat{F})$ against the number of selected features in Figure 2 and Figure 3, respectively. SPEC performs very poorly on high-dimensional datasets. We observe clearly a trend that IVFS keeps lifting its performance as number of features increases. This robustness comes from the fact that the inclusion value intrinsically contains rich information about features interactions.

Stability. We bootstrap samples from original dataset to mimic the process of sampling from population. Denote $\hat{F}_X$ and $\hat{F}_B$ the subset chosen based on original data and bootstrap data respectively. We count $|G|$ with $G = \{ f : f \in \hat{F}_X, f \notin \hat{F}_B \}$. We use same parameters as for testing the running time. The results are averaged over 5 repetitions. In principle, the selected feature pool should not vary significantly if we only change a few samples (by bootstrap), since the “truly” important features should be independent of the samples. In Table 2, we see that IVFS-$l_\infty$ is hardly affected by the bootstrapping process, while other methods are much more sensitive and gives very different solutions.

Efficiency-capacity trade-off. From Table 1, we also see that IVFS-$l_\infty$ is comparable with MCFS in terms of running time, and is more efficient than GLSPFS. In particular, GLSPFS may run very slowly when it takes a long time to converge (e.g., SMK-CAN-187). Additionally, for datasets with large size, SPEC and GLSPFS gets slower due to large matrix inverse and singular value decomposition.

The computational cost of our IVFS algorithms depends tightly on the sub-sampling rate $n/n$ and number of random subsets $k$. As one would expect, there exists a trade-off between computational efficiency and distance preserv-
Table 2: Stability under bootstrap: the number of different selected features between original data and bootstrap data.

| Dataset       | SPEC | MCFS | GLSPFS | IVFS-L∞ | 1000 |
|---------------|------|------|--------|---------|------|
| CLL-SUB-111   | 187.2| 271.8| 243.6  | 12.2    |      |
| Lymphoma      | 156.6| 260.8| 260.4  | 4.4     |      |
| Orlraws10P    | 119.2| 283.6| 267.2  | 7.8     |      |
| Pixrow10P     | 143.0| 284.8| 266.0  | 6.8     |      |
| Prostate-GE   | 125.8| 264.4| 236.2  | 12.0    |      |
| SMK-CAN-187   | 125.8| 264.4| 236.2  | 18.6    |      |
| WarpPIE10P    | 56.0 | 241.6| 191.0  | 4.8     |      |
| COIL20        | 14.0 | 178.2| 115.4  | 5.6     |      |
| Isolet        | 12.8 | 124.5| 80.0   | 3.3     |      |
| RELATHE       | 21.4 | 145.0| 102.8  | 8.4     |      |

Note that in our experiment, when the sample size is relatively large (i.e., greater than 1000), we simply fix \( \hat{n} = 100 \), and IVFS still outperforms other methods (on COIL20, RELATHE). Thus, to achieve better efficiency, we recommend practitioners to set \( k = 1000 \) as default. For \( \hat{n} \), if the data size is not very large, we suggest using \( \hat{n} = 0.1n \) as the first try. Otherwise, one may threshold \( \hat{n} \) at a small number (e.g., 100 in our experiment). This way, the running time, when fixing \( k \) and \( \hat{d} \), becomes approximately constant.

Figure 4: Performance comparison across different number of subsets \( k \) and sub-sampling ratio \( \hat{n}/n \).

Figure 2: \( L^2 \) norm vs. number of features.

Figure 3: Bottleneck distance vs. number of features.

Discussion

For supervised random forest, it is conventional to set as default \( d = \sqrt{d} \), which is around \( 10^2 \) when \( d \) is around \( 10^4 \). However, in this case we have to use large \( k \) to make sure that every feature is evaluated for not too few times. We observe that there is not much gain in performance with small \( d \) and large \( k \) combination. Thus, we suggest to use a relatively large \( d \) to speed up the algorithm for higher efficiency.

It is worth mentioning that there are also hybrid strategies for feature selection, where one first determines a pool of features, and then select ultimate feature set from the pool through another round of screening. One thing we observe from the experiments is that IVFS is much more stable and robust than other algorithms, which means that the features selected are mostly “good” features that help with reducing the loss. Therefore, IVFS is also suitable to be applied in the pool selection procedure for such hybrid methods.

Conclusion

In this paper, we propose IVFS, a unified feature selection scheme based on random subset methods. After we show its connection with existing methods such as random forest and RKNN, we propose IVFS-\( L_\infty \) and several variants that can preserve the pairwise distance and topological signatures (persistent diagram) of the original dataset more precisely than the competing similarity preserving algorithms. This would be very helpful for applications requiring high-level distance preservation, e.g. topological data analysis. In the experiments, we evaluate the distance preserving capability of different algorithms through to demonstrate the effectiveness of the proposed IVFS algorithms. We also demonstrate that a sharp sub-sampling rate can be effectively adopted on this problem to speed up the algorithm on large datasets.
References

[Adams et al. 2017] Adams, H.; Emerson, T.; Kirby, M.; Neville, R.; Peterson, C.; Shipman, P. D.; Chepushanov, S.; Hansen, E. M.; Motta, F. C.; and Ziegelmeier, L. 2017. Persistence images: A stable vector representation of persistent homology. J. Mach. Learn. Res. 18:8.1–8.35.

[Asuncion and Newman 2007] Asuncion, A., and Newman, D. 2007. Uci machine learning repository.

[Breiman 2001] Breiman, L. 2001. Random forests. Machine learning 45(1):5–32.

[Bubenik 2015] Bubenik, P. 2015. Statistical topological data analysis using persistence landscapes. J. Mach. Learn. Res. 16(1):77–102.

[Cai, Zhang, and He 2010] Cai, D.; Zhang, C.; and He, X. 2010. Unsupervised feature selection for multi-cluster data. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 333–342.

[Chazal, De Silva, and Oudot 2014] Chazal, F.; De Silva, V.; and Oudot, S. 2014. Persistence stability for geometric complexes. Geometriae Dedicata 173(1):193–214.

[Du and Shen 2015] Du, L., and Shen, Y. 2015. Unsupervised feature selection with adaptive structure learning. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 209–218.

[Edelsbrunner and Harer 2010] Edelsbrunner, H., and Harer, J. 2010. Computational topology: an introduction. American Mathematical Soc.

[Foroutan and Sklansky 1987] Foroutan, I., and Sklansky, J. 1987. Feature selection for automatic classification of non-gaussian data. IEEE Transactions on Systems, Man, and Cybernetics 17(2):187–198.

[Geurts, Ernst, and Wehenkel 2006] Geurts, P.; Ernst, D.; and Wehenkel, L. 2006. Extremely randomized trees. Machine Learning 63(1):3–42.

[Guo and Zhu 2018] Guo, J., and Zhu, W. 2018. Dependence guided unsupervised feature selection. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), 2232–2239.

[Guyon and Elisseeff 2003] Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3(Mar):1157–1182.

[He, Cai, and Niyogi 2006] He, X.; Cai, D.; and Niyogi, P. 2006. Laplacian score for feature selection. In Advances in Neural Information Processing Systems (NIPS), 507–514.

[Hofer et al. 2017] Hofer, C.; Kwitt, R.; Niethammer, M.; and Uhl, A. 2017. Deep learning with topological signatures. In Advances in Neural Information Processing Systems (NIPS), 1634–1644.

[Li and Zhang 2017] Li, P., and Zhang, C.-H. 2017. Theory of the GMM kernel. In Proceedings of the 26th International Conference on World Wide Web (WWW), 1053–1062.

[Li et al. 2012] Li, Z.; Yang, Y.; Liu, J.; Zhou, X.; and Lu, H. 2012. Unsupervised feature selection using nonnegative spectral analysis. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI).

[Li et al. 2018a] Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; and Liu, H. 2018a. Feature selection: A data perspective. ACM Computing Surveys (CSUR) 50(6):94.

[Li et al. 2018b] Li, J.; Wu, L.; Dani, H.; and Liu, H. 2018b. Unsupervised personalized feature selection. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), 3514–3521.

[Li, Harner, and Adjeroh 2011] Li, S.; Harner, E. J.; and Adjeroh, D. A. 2011. Random kNN feature selection—a fast and stable alternative to random forests. BMC bioinformatics 12(1):450.

[Li et al. 2018c] Li, J.; Wu, L.; Dani, H.; and Liu, H. 2018c. Stable alternative to random forests. In Advances in Neural Information Processing Systems (NeurIPS), 2987–2997.

[Yu and Liu 2003] Yu, L., and Liu, H. 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the Twentieth International Conference on Machine Learning (ICML), 856–863.
[Zhao and Liu 2007] Zhao, Z., and Liu, H. 2007. Spectral feature selection for supervised and unsupervised learning. In Proceedings of the Twenty-Fourth international conference on Machine learning (ICML), 1151–1157.

[Zhao et al. 2013] Zhao, Z.; Wang, L.; Liu, H.; and Ye, J. 2013. On similarity preserving feature selection. IEEE Transactions on Knowledge and Data Engineering 25(3):619–632.

[Zhu et al. 2017] Zhu, X.; Zhu, Y.; Zhang, S.; Hu, R.; and He, W. 2017. Adaptive hypergraph learning for unsupervised feature selection. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI), 3581–3587.