2020 IEEE International Conference on Big Data (Big Data)

Interpretable feature subset selection: A Shapley value based approach

Sandhya Tripathi
Department of Anesthesiology
Washington University at St. Louis
sandhyat@wustl.edu

N Hemachandra
IE&OR
Indian Institute of Technology Bombay
nh@iitb.ac.in

Prashant Trivedi
IE&OR
Indian Institute of Technology Bombay
trivedi.prashant15@iitb.ac.in

***Abstract*—While performing Feature Subset Selection (FSS) to identify important features, a weight is assigned to each feature that is not necessarily meaningful or interpretable w.r.t. final task and in turn leads to non-actionable information. To provide a solution to this problem of interpretable FSS, we introduce a novel notion of classification game with features as players and hinge loss based characteristic function. We use the Shapley value of this game to apportion the total training error to explicitly compute the contribution of each feature (Shapley Value based Error Apportioning, SVEA) to the total training error. We formalize the notion of interpretability in FSS by identifying 3 final task related conditions. We empirically demonstrate that features with SVEA values less than zero are the dominant ones; this set is unique for a dataset as Shapley value is unique for a game instance. For the datasets that had negative apportioning, we observe a high value of the power of classification, P_{SV}. It compares the performance of a set of linear and non-linear classifiers learned on Shapley value-based important features and the full feature set, in most of the cases. We customize a known Monte Carlo based approximation algorithm to avoid expensive Shapley value computations. We demonstrate the sample bias robustness of SVEA scheme by providing interval estimates. We illustrate all the above aspects on both synthetic and real datasets and showed that our scheme out-performs many existing approaches like recursive feature elimination and ReliefF in most of the cases.**

1. Introduction

"What is the guarantee that a given model uses important and relevant features among the given features?" This question has been the topic of research for decades in many learning areas, including supervised learning. To address this question in a binary classification task, we present a cooperative game-theoretic framework for feature subset selection. We introduce a classification game with features as players and hinge loss based characteristic function (in terms of Linear Programs, LPs). Since the classifier that doesn't involve any features has non-zero training error, the challenge in defining a cost game is to deal with the requirement that characteristic function's value should be zero for the empty coalition.

We overcome this challenge by suitably defining a value game and apportioning the total training error of the hinge loss based linear classifiers using an affine transformation of the Shapley value of the value game. As Shapley value allocates the total training error to each feature based on its proportional contribution ('paid as per your participation, no more, no less'), it is theoretically sound and has been famous as a cost allocation measure [1], [2], [3] and in other areas as well [4]. It also captures the interactions among features by the marginal contribution of each feature in a fixed group of features. Thus, it is a suitable choice for tasks like Feature Subset Selection (FSS).

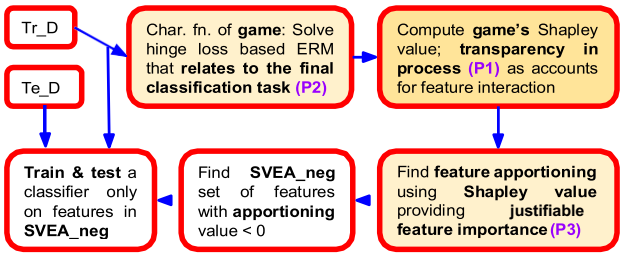


Figure 1: Flow chart describing interpretable FSS scheme SVEA with highlighted properties, P1, P2, and P3 defined in Def. 1. Here, Tr D and Te D denote train and test dataset.

In the context of classification game, the additivity axiom of Shapley value [5], [6], [7] requires that the allocation of total training error by combining two data sets is equal to the sum of the allocations from the different datasets; this is not possible for two distinct data sets. To circumvent this problem, we have used Young's strong monotonicity based axiomatic approach [8] that bypasses the requirement of additivity axiom. We also note that the use of Shapley value in other contexts, such as explaining a prediction is criticized by many researchers in different ways [9]; it was pointed out that additivity axiom need not hold in such applications of Shapley value.

Our **major contributions** are:

(1) Identification of features whose joint contribution to label prediction is substantial (Section 3.1). There is a universal threshold of 0 on Shapley value-based error ap-

. Detailed version of paper available at https://arxiv.org/abs/2001.03956

978-1-7281-6251-5/20/$31.00 ©2020 IEEE

portioning (SVEA) values in our scheme to select feature subset for all datasets. Hence, our scheme doesn't require a user to choose a threshold either on feature importance value or size of feature subset.
(2) Using a mix of linear and non-linear classifiers on datasets ranging from low to moderately sized feature set and examples, we measure the effectiveness of SVEA by $P_{SV}(K)$, the power of classification of SVEA based subset K. We observed that in most of the datasets, this value is approximately 0.90 except for Thyroid dataset (Section 4). In Section 3.2, we provide t-distribution based confidence intervals to show unique SVEA's sample bias robustness.
(3) We explicitly define the notion of an interpretable FSS scheme (Definition 1) and evaluate a range of FSS scheme w.r.t. the proposed definition. We observe that our scheme satisfies all conditions required of an interpretable FSS scheme.

Note that, if one needs the top l features out of n ($l \leq n$), then based on this requirement, one can rank the features by their SVEA values and identify the l-sized subset. Also, if there is a user-given threshold say χ, that is other than 0, the SVEA scheme can also identify the feature subset corresponding to this threshold. Feature subset based on SVEA is unique as Shapley value is unique for a dataset.

In addition to the above-listed contributions, we present a sampling-based approximation algorithm built on [10] that does not require computing characteristic function (LP) for the 2^n subset of features all at once; instead, compute it only when a particular subset of features is sampled (Appendix B). We also considered another variant where the linear classifier based training error is regularized and computationally observed that the feature subset selected is the same as that of the unregularized model (Section 3.3).

Our idea that thresholding the modified Shapley value of classification game at 0 identifies the features with substantial joint contribution to the prediction has following motivation. Suppose among a group of players (features), one player has sufficient resources so that it has the power to work (classify) alone. Let us call it a dominant player. Now, if the other players (features) ask this dominant player to join their coalition (to form a classifier), then it asks them for a payoff. Since, the quantity to be divided is an error (cost), for such dominant players, the payoff is in the form of modified Shapley value being negative. We demonstrate this phenomenon in Pima Diabetes dataset where, knowing the blood sugar level (feature) is sufficient to decide whether the patient has diabetes or not.

An innate understanding of how our SVEA scheme possesses explainability and interpretability (formally in Definition 1) is as follows. Explainability in our scheme refers to its ability to provide a reason for selecting a feature as important using its SVEA value; a feature with negative SVEA lowers the total training error. An important feature subset constituting such features makes FSS (using the SVEA scheme) explainable. Interpretability in the context of the SVEA scheme includes accounting for possible interactions among features using Shapley value, using the training error similar to the one used in final classification task and mapping SVEA (importance) value of a particular feature to an apportioning of training error by the Shapley value of the well defined classification game. The SVEA values can either be negative or positive; features with negative value can be interpreted as the dominant ones (more details in Section 3.1).
Organization Section 1.1 provides details about where our work stands w.r.t. existing literature. The binary classification game model is formulated in Section 2. Next, in Section 3, we present the main insights about interpretable FSS. In Section 4, we demonstrate empirical evidence in support of the proposed scheme. We conclude the paper in Section 5.

1.1. Related work

In this section, we review some existing work on feature subset selection problem. We will describe the use of Cooperative Game Theory (CGT) in FSS and the interpretability aspects of the FSS methods.

Based on the search strategy used, [11] classifies the feature subset selection techniques into three categories, viz., filter techniques, wrapper techniques, and embedded methods. Recursive feature elimination by [12] and ReliefF by [13] are the most popular wrapper and filter methods respectively. Another approach by [14] proposes a graph-theoretic clustering-based FSS scheme that first clusters the features and then choose a representative from each cluster to get the final important feature set. An interesting idea of instance dependent FSS for a general task (classification or regression) is presented in [15], where the authors compute saliency for each feature by identifying a task and loss dependent gain function.

CGT in feature selection: In [16] authors have proposed a contribution selection algorithm that uses Shapley value to improve upon wrapper techniques like backward elimination and forward selection. The solution concepts such as Shapley value and Banzhaf index are used by [17] and [18] respectively to compute the importance of features which is further used with the filter methods based on information-theoretic ranking criteria. A good amount of work by [19], [20], [21], [22] also uses Shapley value for feature importance while dealing with medical data. However, the definition of payoff function is not explicit and the algorithms depend on user given parameters. All the methods mentioned above use CGT mainly to give additional information to either a wrapper or filter method. However, CGT is central to our scheme as it uses an affine transformation of Shapley value of the classification game which further provides interpretability and explainability to the selected feature subset. Also, the existing methods require a user given threshold on the contribution value, whereas, for us, the threshold of 0 (a universal threshold) is used to select the feature subset.

Interpretability in feature subset selection: Feature subset selection being an integral part of any learning model needs interpretable and explainable methods too. A visual explanation and interpretation approach for dimension reduction is presented by [23]. Mutual information based

feature selection method that uses the unique relevant information and show its importance in health data is given in [24]. Local information based interpretable feature subset selection is also studied by [25]. Another more recent approach called Informative Variable Identifier (IVI) in FSS by ensemble category is proposed by [26] where they also provide levels of interpretability in FSS. However, their scheme relies on statistical properties of feature distribution to incorporate feature interactions. We provide a single definition for an FSS algorithm to be interpretable in Definition 1. FSS methods that mitigate the bias amplification in linear models have been proposed by [27] wherein the authors have presented two new feature selection algorithms for mitigating bias amplification in linear models, and show how they can be adapted to convolutional neural networks efficiently. The influence function is used to remove the features which have bias towards the prediction. However, our focus is on the feature subset selection. We have also addressed the issue of sample biasedness, but that is different from the biasness of the features towards the prediction.

Shapley value for explaining a prediction and data valuation: CGT has been used in literature either for explaining a model's prediction [28], [29], [30] or for data valuation by using the Shapley value [31], [32]. The difference between our work of using CGT for FSS (*before training*) and existing work using it for explaining predictions (*after training*), has been also clarified and highlighted by [33], where they treat feature importance across all the training data and attribution (explaining model prediction) as two separate problems. CGT based data valuation work focuses on selection of the most relevant data points to apportion the overall profit among various contributors by considering data points as players, unlike our work, where we model features as players. We would like to emphasize that our scheme is not just a global version of Shap [30] as the later scheme averages over the Shap values for every feature across data points to get a summary importance. Instead, we have an explicit game formulation (whose relevance has been already pointed out by [34] in a different setup of explaining predictions) whose Shapley values are used as feature contribution to the training error. Hence, our definition is novel and takes natural approach of Empirical Risk Minimization (ERM).

2. Training error based classification game

In this section, we will describe definitions and notations to be used throughout the paper. The training error incurred by a subset of feature is then introduced, which is used to define the classification game – a cooperative game which is a central and novel contribution of our work.

2.1. Notations and Preliminaries

In this section, we introduce some cooperative game [7], [35] and classification [36], [37] terminology and concepts to provide a better understanding of the connection which we will be studying in rest of the paper.

Cooperative game theory [7], [35]: The Transferable Utility (TU) cooperative game is a pair (N, v) where $N = \{1, \ldots, n\}$ is a set of players and $v : 2^N \mapsto \mathbb{R}$ is the characteristic function, with $v(\emptyset) = 0$. Any subset $S \subseteq N$ of player set is called the coalition of players in set S. Set of all players is referred to as grand coalition. One of the major problem in CGT is the allocation of the payoff of grand coalition among all the players. There are various axiomatic approaches, but we are using Young's axiomatic approach [8]. According to Young's axiomatic approach, Shapley value [7], [35] is a unique, symmetric, and strongly monotonic solution concept defined as a mapping $\phi : \mathbb{R}^{2^n-1} \mapsto \mathbb{R}^n$ where $\forall j \in N, \ \ \forall\ v \in \mathbb{R}^{2^n-1}$ we have:

$$\phi_j(v) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(n-|S|-1)!}{n!}[v(S \cup \{j\}) - v(S)].$$

Classification setup [36], [37]: Let $\mathcal{X}$ be the feature space and $\mathcal{Y}$ be the label set. Let $\mathcal{D}$ be the joint distribution over $\mathbf{X} \times Y$ with $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^n$ and $Y \in \mathcal{Y} = \{-1, 1\}$. $\mathbf{X}$ can include categorical features after suitably prepossessing them either by using dummy encoding for nominal ones or by using ordered numbers for ordinal variables. Let the decision function be $f : \mathbf{X} \mapsto \mathbb{R}$ and hypothesis class of all measurable functions be $\mathcal{H}$. Let the linear hypothesis class be $\mathcal{H}_{lin} = \{(\mathbf{w}, b), \mathbf{w} \in \mathbb{R}^n, b \in \mathbf{R}\}$. In the game construction, we restrict the hypothesis class to be $\mathcal{H}_{lin}$ because Shapley value's underlying assumption that requires the formation of grand coalition is violated by use of non-linear classifiers while defining the characteristic function. However, for the final classification task, we use both linear and non-linear classifiers (Section 4). We have an i.i.d. sample of size m from distribution $\mathcal{D}$, viz., $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ is the value of the feature and $y_i \in \{-1, 1\}$ is the label for i^{th} data point. We use hinge loss based ERM setup because in addition to many desirable properties such as classification calibration and the large margin it imparts to classifiers, it leads to an LP which can be solved in polynomial time.

2.2. Training error function

Given the dataset/sample $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ with features $N = \{1, \ldots, n\}$, we consider a training error function, $tr\_er(S, m)$ associated with all possible subsets $S \subseteq N$ when sample size is m. We define $tr\_er(\emptyset, m)$ as hinge loss based training error of an intercept only classifier and denote it by $\tilde{c}(m) := tr\_er(\emptyset, m)$.

$$\begin{aligned} tr\_er(\emptyset, m) = & \min_{b, \{\xi_i\}_{i=1}^m} \frac{1}{m} \sum_{i=1}^m \xi_i \\ \text{s.t. } & y_i b \geq 1 - \xi_i, \quad \forall i = 1, \ldots, m \\ & \xi_i \geq 0, \quad \forall i = 1, \ldots, m. \end{aligned} \tag{1}$$

Similarly, we define the training error, $tr\_er(S, m)$ for any nonempty subset $S = \{j_1, j_2, \ldots, j_r\}$ of size r with r distinct elements/features. This would be minimal hinge loss of the classifier $(w_{j_1}^*, \ldots, w_{j_r}^*, b_r^*)$ obtained from the dataset

projected to r-dimensional subspace, i.e., dataset having feature values $\{x_{ij_1}, \ldots, x_{ij_r}\}_{i=1}^m$ and label $\{y_i\}_{i=1}^m$.

$$\begin{aligned} tr\_er(S,m) = & \min_{w_{j_1},\ldots,w_{j_r},b_r,\{\xi_i\}_{i=1}^m} \frac{1}{m}\sum_{i=1}^{m}\xi_i \\ \text{s.t. } & y_i\left(\sum_{j\in S} w_j x_{ij} + b_r\right) \geq 1-\xi_i,\ \forall i \in [m] \\ & \xi_i \geq 0, \quad \forall i = 1,\ldots,m. \end{aligned} \tag{2}$$

When $S = N$, we have $tr\_er(N,m)$, i.e., the minimal hinge loss based empirical risk of the classifier $(\mathbf{w}_N^*, b_N^*)$ when the given dataset is n dimensional (all n feature values from the sample D are used). Note that the variables used in each ERM are local to that optimization problem only.

As conventional cooperative games assume $v(\emptyset) = 0$, training error function $tr\_er(\cdot, m)$ with $tr\_er(\emptyset, m) \neq 0$ cannot be a valid characteristic function. To circumvent this problem, we define a payoff/value game with characteristic function $v(S,m)$[1] given below:

$$v(S,m) = tr\_er(\emptyset,m) - tr\_er(S,m),\ \forall\ S \subseteq N. \tag{3}$$

$v(S,m)$ represents the marginal improvement in the training error obtained due to the presence of the features in S. As, $v(\emptyset, m) = 0$, it is a valid characteristic function also. This characteristic function along with the feature set N defines a TU **classification game** $(N, v(\cdot,m))$. Further, the characteristic function $v(S,m)$ is monotonic w.r.t. the coalitions, which is an important property from the perspective of allocation. This property is formalized in Proposition 1 with the proof being available in Appendix A.1.

Proposition 1. If $(N, v(\cdot,m))$ is a classification game, then the characteristic function $v(\cdot,m)$ is monotonic, i.e., $\forall\ \ S \subseteq T \subseteq N, \quad v(S,m) \leq v(T,m)$.

2.3. Training error allocation using Shapley value

As the Shapley value solution concept has the idea of allocation based on a feature's marginal contribution (no more, no less), it emerges as a suitable candidate for apportioning of $v(N,m)$ among the features in a classification game and its Shapley values for a feature $j \in N$, is:

$$\phi_j(N, v(\cdot,m)) = \sum_{S\subseteq N\setminus\{j\}} \frac{|S|!(n-|S|-1)!}{n!}[v(S\cup\{j\},m) - v(S,m)]. \tag{4}$$

Using this Shapley value, Theorem 1 provides an equitable training error allocation among features. We refer to it as **Shapley value based error apportioning (SVEA)** denoted by $e_j(tr\_er(N,m)),\ \forall\ j \in N$; as we see below, it is an affine transformation of Shapley value for feature $j \in N$. A proof of Theorem 1 is available in Appendix A.2. As Shapley value is unique for an instance of a game which in our case is a sample, SVEA is also unique for a sample from a dataset.

1. Characteristic function as defined here depends on sample size m, so we use m as an argument in $v(\cdot,m)$. We later use sub-samples to avoid sample bias.

Theorem 1. The unique Shapley value based error apportioning, $e : \mathbb{R}^{2^n-1} \to \mathbb{R}^n$ of the total training error, $tr\_er(N,m)$ among all the features is given by

$$e_j(tr\_er(N,m)) = \frac{\tilde{c}(m)}{n} - \phi_j(N, v(\cdot,m)), \quad \forall j\ \in N. \tag{5}$$

For notational convenience, hereafter, we will denote the allocation of training error to feature j, by $e_j(m)$ and Shapley value of feature j by $\phi_j(m)$for sample size of m.

In general, the problem of computing Shapley value is known to be NP-hard [38]. Also, it has high space complexity due to the space requirement of storing $n!$ permutations or $2^n - 1$ characteristic functions. To bypass this issue, we adapt the approximation algorithm given by [10] for computing the Shapley value of features in the classification game $(N, v(\cdot,m))$. The advantage of using this algorithm is that the characteristic function is calculated for a coalition as and when required in the marginal contribution sum. Note that the computation of $tr\_er(S,m)$ for a coalition S is scalable as it is by an LP. Algorithm and related details are available in Appendix B. In Section 4, we use this approximation for datasets with $n \geq 10$. To evaluate the quality of Shapley value estimates, we computed their difference from the true Shapley value for datasets with $n < 10$ and observed that use of 100 Monte Carlo (MC) samples lead to a min 0.5 % and max 10% error over ten trials (different train and test partitioning) across all datasets. If the MC samples are increased to 1000, this error comes down to a max 4 percent. In our preliminary experiments, as the sign of apportioning and ordering of features doesn't change with the use of 100 MC samples and 1000 MC samples, we stick to using 100 MC samples only. Now we will address some interpretability aspects of our method.

3. Interpretable feature subset selection

Based on the apportioning of the total training error, $tr\_er(N,m)$, among the features, we are deciding whether the feature is important or not (details available in the subsequent subsection). One of the key points to note is that for some features $j \in N$, $e_j(m) < 0$. The intuition is as follows: suppose a player (feature) is so dominant that it can work (classify) alone. Now, if the other players (features) ask this dominant player to join their coalition (to form a classifier), then it asks them for a payoff. Since the quantity to be divided is an error (cost), for such dominant players, the payoff is in the form of SVEA being negative. We formally present this idea in Proposition 2 for the two-player case. The proof is available in Appendix A.3.

Proposition 2. Consider a 2-feature classification game $(N, v(\cdot,m))$ with training error function $tr\_er(\{1\},m) = q > 0$, $tr\_er(\{2\},m) = Q > 0$, $tr\_er(\{1,2\},m) = q' \leq \min\{q, Q\}$. If $\frac{Q}{2} > q$, then SVEA of $tr\_er(\{1,2\},m)$ is such that $e_1(m) \leq 0$ and $e_2(m) \geq 0$.

Generalizing the notion of Proposition 2 for n players, the apportioning $\{e_j(m)\}_{j\in N}$ of $tr\_er(N,m)$ can provide

us with various insights. Based on the above arguments, in this work, we study the role of those features for which SVEA is negative in feature subset selection. In particular, we show SVEA based decision of whether a feature is to be selected or not is easily interpretable. We further show that these SVEA values being less than zero is not an artifact of the sample, but dataset property and hence are robust to potential sample bias. To formalize the notion of interpretabilty in FSS, we define it as follows:

Definition 1. A scheme for FSS is said to be interpretable if it satisfies following conditions:

(P1) **Transparency in the process:** The process of finding the feature importance should be transparent and it should be clear how the feature interactions are being accounted for.

(P2) **Relation to final task:** The feature importance computation should be based on a criterion which takes into account their role in final task (classification).

(P3) **Justifiable importance values:** Feature importance values should have a meaning/justification in the context of the final task in addition to being just called feature contributions.

SVEA is interpretable as it satisfies all above conditions. It accounts for all possible interactions among features using Shapley value (P1), uses the training error similar to the one used in final classification task (P2) and SVEA (importance) value of a particular feature correspond to an apportioning of training error by the Shapley value of the *well defined* classification game (P3). Figure 1 depicts the steps of SVEA scheme in which the above properties are satisfied. Next, we summarize which FSS methods satisfy the different conditions from Definition 1 in Table 1.

Remark: The above definition of interpretablilty is fairly generic, as it can be adopted for other schemes (other than FSS) in a broader task (other than classifier design).

Methods/Conditions	P1	P2	P3
ReliefF [13]	✓	✓	×
RFECV [12]	✓	✓	×
BanzhafI based [18]	✓	✓	×
ShapleyV based [17]	✓	✓	×
LFS [25]	✓	✓	×
IVI [26]	✓	✓	×
SVEA based FSS [26]	✓	✓	✓

TABLE 1: Table summarizing FSS methods w.r.t. interpretability from Definition 1.

3.1. Negative valued SVEA and FSS

We observed that the features for which SVEA is negative (set $SVEA_{neg}$) are the ones whose joint contribution in label prediction is substantial. To formalize this idea, we introduce the notion of the power of classification of a subset, say K, of features, defined below:

Definition 2 (Power of classification of feature subset K, $P_{SV}(K)$). Given a training dataset D of size m with feature values $\{x_{i1}, \ldots, x_{in}\}_{i=1}^{m}$ and labels $\{y_i\}_{i=1}^{m}$, the power of classification of a set of features $K = \{j_1, j_2, \ldots, j_k\} \subseteq N$ is defined as follows:

$$P_{SV}(K) = \frac{\sum_{i=1}^{m_{te}} \mathbf{1}_{[y_i f_K^*(x_{ij_1}, x_{ij_2}, \ldots, x_{ij_k}) \geq 0]}}{\sum_{i=1}^{m_{te}} \mathbf{1}_{[y_i f_N^*(x_{i1}, x_{i2}, \ldots, x_{in}) \geq 0]}}, \quad (6)$$

where $f_K^*(\cdot)$ and $f_N^*(\cdot)$ are the optimal linear classifiers in the respective subspaces, m_{te} (different from D) is the number of sample points used for testing, and $\mathbf{1}_{[A]}$ is the indicator function with value 1 if A holds and 0 otherwise.

The higher the value of $P_{SV}(K)$, the higher the joint influence of the subset K in classification. The powerful subset $K = SVEA_{neg}$ is not pre-decided but determined by SVEA. Due to Shapley value's property of identifying the important players based on their contributions, the SVEA scheme identifies features that play a dominating role in the task of classification and forms a set K. We give details about the FSS interpretation for UCI datasets with the SVEA, $\{e_j(m)\}_{j \in N}$ being negative in Section 4. Besides,in our preliminary experiments we observed that l_1-regularized squared hinge loss based ERM doesn't identify important features for UCI datasets like Heart, Pima, and Thyroid.

3.2. Sample bias robustness of SVEA scheme

To be robust to sample bias, we provide interval estimates for SVEA of features by using multiple sub-samples from a given dataset. A feature's joint contribution in label prediction is substantially high if the interval estimate of SVEA for a feature lies on the left of origin on $\mathbb{R}$. The procedure is first to partition the dataset into multiple disjoint subsamples and compute the apportioning for each sub-sample. Then, a group of 30 such sub-samples is selected and using CLT, the average apportioning $\bar{e}_j^g, j \in N$ for each group g is asymptotically normally distributed with unknown mean μ_e and variance σ_e^2. Next, using $\bar{e}_j^g$, we compute t-distribution based $100(1-\alpha)$ confidence intervals. Let G be the number of groups; then by the definition of confidence intervals, we have following high probability statement:

$$P(e_j^p \in [\bar{\bar{e}}_j \pm t^*_{\alpha/2, G-1}(s_j/\sqrt{G})]) \geq 1 - \alpha, \quad \forall j \in N,$$

where e_j^p is the population mean for the error apportioning value of feature $j \in N$, $\bar{\bar{e}}_j = \frac{1}{G}\sum_g \bar{e}_j^g$ and $s_j = (\frac{1}{G-1}\sum_g (\bar{e}_j^g - \bar{\bar{e}}_j)^2)^{1/2}$ and $t^*_{\alpha/2,G-1}$ is the upper $\alpha/2$ critical value for the t distribution with $G-1$ degrees of freedom. Based on our experiments in Section 4, we observe that the interval estimates also lead to the same threshold of 0 while performing FSS. Also, the conclusions are robust to sample bias due to multiple averaging; thus, the behavior of features with negative SVEA mentioned in Section 3.1 is a property of the dataset and not of a particular sample. The method presented above is tailor-made for the SVEA scheme. A more general framework to address the

instability issue, i.e., change in sample leading to change in feature subset is presented in [39]. The authors first show that any existing stability measure doesn't possess all five desirable properties which a stability measure should have. Then, taking a statistical approach, they propose a novel measure that is treated as an estimator of true stability.

3.3. Characteristic function with regularization

In this section, we first shed some light on the importance of Proposition 1 while using Shapley value for identifying important features in a non-regularized linear classifier hinge loss based ERM problem. Shapley value implicitly assumes that the grand coalition is formed. Monotonicity, i.e., $v(N) \geq v(S), \forall S \subseteq N$ of the characteristic function ensures the formation of grand coalition. To this end, Proposition 1 for the classification game with a linear classifier provides a sufficient condition for the formation of a grand coalition and hence the use of Shapley value to apportion the total value is justified. However, this might not be true for other characteristic function as given below.

To address the issues such as over-fitting, we considered a l_2-regularized version of $tr\_er(S,m)$ defined in Section 2.2. For trade-off parameter $C > 0$,

$$\begin{aligned} tr\_er(S,m) = & \min_{w_{j_1},\ldots,w_{j_r},b_r,\{\xi_i\}_{i=1}^m} C\sum_{i=1}^{m}\xi_i + \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t. } & y_i\left(\sum_{j\in S} w_j x_{ij} + b_r\right) \geq 1-\xi_i, \quad \forall i = 1,\ldots,m \\ & \xi_i \geq 0, \quad \forall i = 1,\ldots,m. \end{aligned} \tag{7}$$

Clearly, using Eq. (7) to compute $v(S,m) = tr\_er(\emptyset,m) - tr\_er(S,m)$ with $tr\_er(\emptyset,m)$ as in Section 2.2 can lead to negative $v(S,m)$. To avoid this issue, we define $v_{reg}(S,m) = tr\_er(\emptyset,m) - \frac{1}{m}\sum_{i=1}^{m}\xi_i^*$ where $\xi_i^*, i = 1,\cdots,m$ is optimal solution of problem in Eq. (7). Even though $v_{reg}(S,m)$ is not shown to be theoretically positive, we observed it to be positive in all our experiments.

We computed SVEA using $v_{reg}(S,m)$ for various real and synthetic datasets across five trials. We tuned the parameter C in the set $\{0.1, 1, 50, 500\}$ for the optimization problem in Eq. (7) when $S = N$ and used the best value of C obtained with $S = N$ in the optimization problem for all other subsets $S \neq N$. We observed that the important feature subset corresponding to those features that have SVEA $e_j(m) < 0$ is the same irrespective of the fact whether regularization is used or not in the characteristic function. This is verified across five trials on the datasets for which the Shapley value can be computed exactly. Details available in Table 2. For datasets where Shapley value approximation algorithm from Appendix B is used, we observe that the subset $SVEA_{neg}$ varies across trials and is different with and without regularization. We repeated this experiment many times and observed different elements in $SVEA_{neg}$. This phenomenon is possibly not the effect of regularization, but that of permutation sampling used while computing Shapley value estimates. To summarize, even though the use of Shapley value in this case could be justified empirically only, its performance is the same as that of linear unregularized and hence regularization in the SVEA scheme is not helpful for feature subset selection.

Dataset (m,n)	$SVEA_{neg}$ (without reg)	$SVEA_{neg}$ (with reg)
sdA (3000,5)	$\{1,4\}$	$\{1,4\}$
sdB (9000,6)	$\{1,3,6\}$	$\{1,3,6\}$
Thyroid (215,5)	$\{4\}$	$\{4\}$
Pima (768,8)	$\{2\}$	$\{2\}$
Heart (270,13)	$\{9,12,13\}$; $\{3,11,12,13\}$; $\{3,9,12,13\}$; $\{3,12,13\}$; $\{3,9,12,13\}$	$\{9,11,12,13\}$; $\{3,12,13\}$; $\{3,9,12,13\}$; $\{3,12,13\}$; $\{3,12,13\}$

TABLE 2: Comparison of important feature subset $SVEA_{neg}$ when the characteristic function was defined with and without regularization over five different trials (train-test partitioning). For $n < 10$, Shapley value is computed exactly and $SVEA_{neg}$ is same. For datasets with $n \geq 10$, the use of Shapley value estimates led to a difference in the sets obtained with and without regularization.

4. Computational experiments

In this section, we empirically demonstrate the implications of SVEA being negative for some features on real-world datasets from [40], [41]. For the FSS interpretation, we train a classifier using SVM (linear and rbf kernel), Logistic Regression (LR), Random Forest (RF) and Multi Layer Perceptron (MLP) to compute P_{SV}. We also compare our SVEA approach to Recursive Feature Elimination with Cross-Validation (RFECV) and ReliefF. Implementation of RFECV with 5-folds was done in Scikit learn module of Python [42]. For ReliefF, we used the implementation of [43] with neighbour parameter $k = 2$. All the above mentioned algorithms are implemented in *Python 3* with *Gurobi 8.0.0* solver for LPs, on a machine equipped with 4 Intel Xeon 2.13 GHz cores and 64 GB RAM. To account for randomness, we repeat each experiment 5 times and report the average test accuracy (and standard deviation).

Real datasets: Demonstration of FSS using $\mathbf{P_{SV}}$ with threshold 0 on SVEA values First, we consider those UCI datasets in which some features have negative SVEA and compute their Power of classification, P_{SV}. Let $SVEA_{neg}$ be the set of features with $e_j(m) < 0$ and $P_{SV}(SVEA_{neg})$ is the power of classification of the set $SVEA_{neg}$. From Table 3, the value of $P_{SV}(SVEA_{neg})$ is close to 1 in most of the cases. Hence, the features in the set $SVEA_{neg}$ have a large joint contribution towards classification. Five different types of classifiers support this behavior, viz., SVM (linear), LR, RF, SVM(rbf) and MLP. Except for the Thyroid dataset, the P_{SV} value for a given dataset is similar across classifiers. Moreover, we have also computed the P_{SV} for all subsets of the $SVEA_{neg}$ set for Heart dataset and observed that in comparison to its subsets, $SVEA_{neg}$ has the highest value

of P_{SV}. For heart dataset, we observed that average accuracy ($\pm$ s.d.) over three trials with only feature 3 is 0.765 ± 0.046, with only feature 12 is 0.697 ± 0.043, with only feature 13 is 0.77 ± 0.453, with features 3 and 12 is 0.771 ± 0.046, with features 12 and 13 is 0.77 ± 0.045 and with features 3, 12 and 13, i.e., $SVEA_{neg}$ set is 0.81 ± 0.008 which is highest among all the subsets.

In addition to the datasets reported in Table 3, we also implemented our SVEA scheme on other large scale datasets like EEG-Eye state dataset (14 features, 14980 examples), Numerai dataset (21 features, 96320 examples). However, we did not observe any feature with negative value of SVEA and hence in Table 3 we only report datasets that had features with SVEA values less than zero.

Real datasets: Demonstration of FSS with a user given threshold and comparison to RFECV and ReliefF For a user given feature size, say l, with due justification in terms of SVEA, our scheme can identify the l sized feature set with best test accuracy based on the ranking of the SVEA values. We demonstrate this property of SVEA based scheme and compare it to RFECV and ReliefF. To do this, we order the features based on the score/SVEA value for each scheme and then plot the SVM test accuracy of linear classifiers learned using first l features (Figure 2). Too few features lead to degradation in performance, and too many features defeat the purpose of feature selection. In comparison to other methods, our scheme achieves the highest accuracy when one looks for a trade-off by selecting a subset of features whose cardinality is neither too small nor too large. Also, if the user given threshold on the number of features is l, then SVEA has the best accuracy as observed in Figure 2 for Magic, Heart, and IJCNN dataset with $l = 2, 3, 5$ respectively.

Using a statistical significance test to compare our scheme to RFECV and ReliefF is not straight forward due to computation of incremental feature accuracy, so we use the measure that given a fixed number of features and a lower bound on the accuracy, a good scheme should identify feature subset leading to high accuracy. However, for the sake of completeness, we still performed many Friedman tests (using Scikit-posthoc package in python) by fixing the number of features across datasets and found no significant difference between the schemes at 5% level of significance for most fixed feature sets. For the cases (fixed feature set) when we observed a significant difference between the accuracy using the Friedman test, we further performed the Nemenyi Posthoc test and observed that SVEA has better accuracy in comparison to ReliefF and RFECV [44].

Real datasets: Sample bias robust interval estimates To address the issue of sample bias while making conclusions based on the SVEA scheme, we will now give some interval estimates of the SVEA estimates for the large datasets. Since the computation of interval estimates require partitioning the whole dataset into many disjoint subsets, large sample-sized datasets are considered. Figure 3 shows t-distribution based 95% confidence intervals of SVEA estimates for synthetic dataset sdB (details in Appendix B) and real datasets Magic, IJCNN [45] and MINIBOONE [46]. There is a partitioning of feature set into two subsets, one in which the features have their SVEA's confidence intervals above the origin and other in which the features have their SVEA's confidence intervals below the origin. As $P_{SV}(SVEA_{neg})$ (given in Table 3) for the latter subset of features is high, one can conclude that these features have a large contribution in label prediction. Since the feature set partitioning is based on interval estimates, the conclusions regarding important features are robust to sample bias.

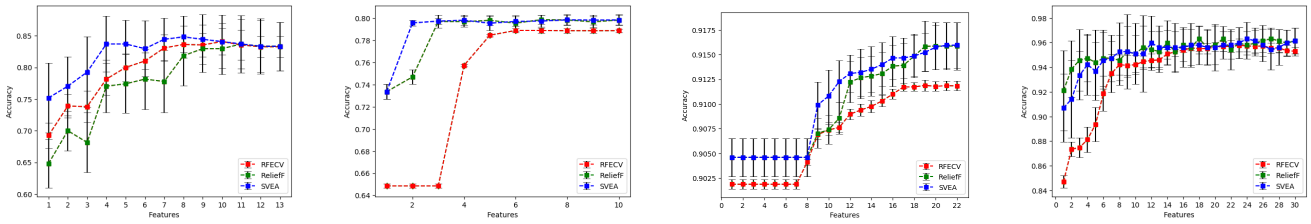
Dataset (m,n)	Clf	Avg Acc ($\pm$std dev)	$SVEA_{neg}$	Avg Acc ($\pm$std dev) $SVEA_{neg}$	P_{SV}
Thyroid (215,5)	SVM	0.89±0.0145	$\{4\}$	0.83±0.0087	0.93
	LR	0.89±0.0315		0.79±0.0328	0.89
	RF	0.93±0.0400		0.76±0.0641	0.82
	SVM_K	0.95+-0.025		0.78+-0.0405	0.82
	MLP	0.94		0.78+-0.0372	0.83
Pima Diabetes (768,8)	SVM	0.77±0.0060	$\{2\}$	0.75±0.008	0.98
	LR	0.75±0.0075		0.73±0.0218	0.97
	RF	0.73±0.0097		0.72±0.0230	0.98
	SVM_K	0.72+-0.0257		0.71+-0.0544	0.99
	MLP	0.67+-0.0269		0.66+-0.0273	0.98
Magic (19020,10)	SVM	0.79±0.0058	$\{9\}$	0.74±0.0060	0.93
	LR	0.79±0.0058		0.73±0.0061	0.92
	RF	0.75 ± 0.0035		0.73± 0.0097	0.97
	SVM_K	0.82+-0.0620		0.84+-0.0544	1.02
	MLP	0.82+-0.007		0.73+-0.0075	0.89
Heart (270,13)	SVM	0.84±0.0478	$\{3, 12, 13\}$	0.81±0.0080	0.96
	LR	0.82±0.0381		0.79±0.0309	0.96
	RF	0.82±0.0343		0.84±0.0578	1.02
	SVM_K	0.64+-0.0608		0.84+-0.0544	1.31
	MLP	0.77+-0.0482		0.79+-0.0035	1.02
IJCNN (126701,22)	SVM	0.91	$\{11, 12, 17, 18, 19\}$	0.90	0.99
	LR	0.91		0.90	0.99
	RF	0.90		0.90	1
	SVM_K	0.979		0.957	0.98
	MLP	0.977		0.956	0.98

TABLE 3: Accuracies of the datasets having negative SVEA for features in $SVEA_{neg}$ for SVM (linear kernel), LR, RF, SVM_K (rbf kernel) and single layer MLP classifier. The sixth column has the accuracy of the classifiers learnt only on features in $SVEA_{neg}$. P_{SV} is the ratio of accuracies in column 3 and column 5 when the important feature subset is $(SVEA_{neg})$. SVM, LR, SVM_K parameter $C \in \{0.1, 1, 50, 500\}$, RF parameters $n\_estimators \in \{0.1, 1, 50, 500\}$ and $max\_depth = 2$, SVM_K parameter $\gamma = \frac{1}{n * Var(X)}$, MLP regularization parameter $\alpha \in \{10^{-7}, \ldots, 10^{-1}\}$, constant learning rate of 10^{-3}, Relu activation and Adam solver. m is the sample size and n is the number of features. No averaging is done for IJCNN as the train-test (35000+91701) partitioning is already available from the source.

5. Discussion and looking ahead

We model a binary classification problem as a cooperative game with features as players and hinge loss based ERM's optimal value as a characteristic function; we introduce the notion of classification game. It accounts for interactions between the features by using a sound solution concept, Shapley value, which uniquely apportions the total training error among the features (SVEA).

SVEA scheme identifies a unique set of features whose joint contribution to prediction is substantial ($SVEA_{neg}$) and doesn't require the user to provide a threshold on importance values or the number of important features. We also



(a) Dataset: Heart(270,13) (b) Dataset: Magic(19020,10) (c) Dataset: IJCNN(126701,22) (d) Dataset: Wdbc(569,30)

Figure 2: Plot of test accuracy vs number of features used to train the linear classifier using SVM. For each scheme, we have 95% error bar computed over 5 iterations. Given a fixed number of features and a lower bound on accuracy, SVEA provides the feature subset which leads to highest test accuracy in most of the cases. Number of trials in the plot for Magic dataset is three; for other datasets number of trials is five.

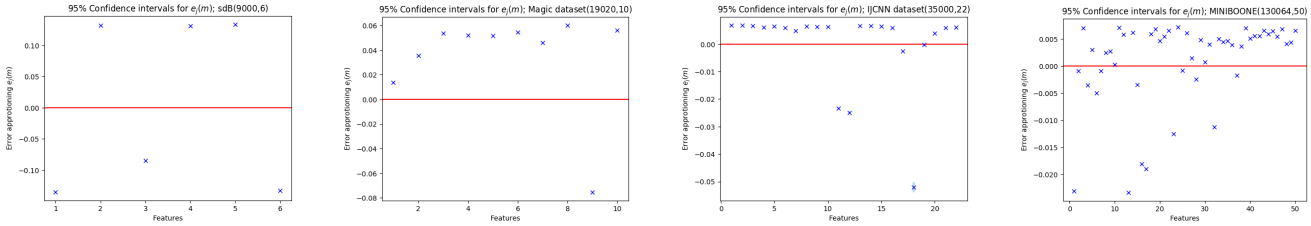


Figure 3: Above plots depict 95% confidence intervals for SVEA of features for 6 dimensional synthetic dataset sdB and 3 UCI datasets. As the importance of a feature is based on intervals of SVEA, we can say that on an average the population value of SVEA would lie inside the interval estimates 95 out of 100 times (which are below 0) and hence important features via SVEA is a pattern manifestation of underlying population.

introduce the notion of interpretable FSS which in addition to providing transparency in the procedure and relation to final prediction, also asks that the feature importance values map to some tangible quantity in the context of final task. SVEA based FSS is interpretable because one can interpret the SVEA value as the error contribution of a feature to the total training error incurred while learning a hinge loss based linear classifier. Also, our approach can provide to the user, a given number of important features by ranking SVEA values. SVEA based FSS is evaluated by computing power index P_{SV}. We implement a mix of linear and non-linear classifiers on UCI datasets to demonstrate that P_{SV} across different classifiers is close to 1. We provide interval estimates for SVEA values so that our FSS is also robust to sample bias. It compares favorably with the existing feature selection schemes RFECV and ReliefF.

We are currently pursuing the use of the SVEA scheme in dimension reduction and in excess 0-1 risk decomposition of a finite sample-based classifier. Also, we believe that SVEA values can be interpreted as estimates of true unknown hinge risk of a feature. A comparison of SVEA from 0-1 loss function and other surrogate loss function based classification games would be interesting to explore; a ranking of surrogate losses for the FSS task can be expected.

Appendix A.

A.1. Proof of Proposition 1

Proof: Consider the optimization problem in Section 2.2 solved to obtain $tr\_er(T,m)$ and $tr\_er(S,m)$, say P_T and P_S for coalitions T and S respectively. Now if $S \subseteq T$, then in addition to the variables in the optimization problem P_S, the optimization problem P_T will have extra variables to solve for. However, a feasible (including optimal) solution in P_S will still remain feasible for P_T by assigning the extra variables a zero value. This implies that minimization in P_T is over a larger feasible set and the objective value of P_T (i.e., $tr\_er(T,m)$) would be upper bounded by the objective value of P_S (i.e., $tr\_er(S,m)$). Therefore, the training error due to features in T will be smaller than that of the training error due to the features that come from all its subset, i.e.,

$$\forall S \subseteq T \subseteq N, \quad tr\_er(T,m) \leq tr\_er(S,m). \tag{8}$$

The result follows by using $tr\_er(\emptyset,m) = \tilde{c}(m) \geq 0$ and the transformation given in Eq. (3). □

A.2. Proof of Theorem 1

Proof: From Eq. (4), the Shapely value for a player j is given by

$$\phi_j(N, v(\cdot,m)) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(n-|S|-1)!}{n!} [v(S \cup \{j\}, m) - v(S,m)].$$

Using efficiency axiom for Shapley value, we have

$$\begin{aligned}
\sum_{j \in N} \phi_j(N, v(\cdot,m)) &= v(N,m) \\
\sum_{j \in N} \phi_j(N, v(\cdot,m)) &= tr_er(\{\emptyset\},m) - tr_er(N,m) \\
\implies \quad tr_er(N,m) &= tr_er(\{\emptyset\},m) - \sum_{j \in N} \phi_j(N, v(\cdot,m)) \\
&= \sum_{j \in N} \left(\frac{\tilde{c}(m)}{n} - \phi_j(N, v(\cdot,m)) \right).
\end{aligned}$$

The last equation follows from the use of LP given in Section 2.2. Hence, the contribution of feature j in the total training error is given as follows:

$$e_j(tr\_er(N,m)) = \frac{\tilde{c}(m)}{n} - \phi_j(N, v(\cdot, m)).$$

□

A.3. Proof of Proposition 2

Proof: From Theorem 1, the apportioning of the total training error is:

$$e_j(m) = \frac{\tilde{c}(m)}{n} - \phi_j(N, v(\cdot, m)) \quad \forall j \in N.$$

Substituting the value of $\phi_j(N, v(\cdot, m))$ in above equation, we have

$$e_j(m) = \frac{tr\_er(\{j\}, m)}{n} + \sum_{\substack{S \subseteq N \setminus \{j\} \\ S \neq \emptyset}} \frac{|S|!(n-|S|-1)!}{n!} [tr\_er(S \cup \{j\}, m) - tr\_er(S, m)].$$

When $n = 2$, we have,

$$\begin{aligned} e_1(m) &= \frac{1}{2}(tr\_er(\{1\}, m) + (tr\_er(\{1,2\}, m) - tr\_er(\{2\}, m)) \\ &\leq \frac{q}{2} + \frac{q' - Q}{2} \leq q - \frac{Q}{2}, \quad (\because q' \leq \min\{q, Q\}) \end{aligned}$$

The last inequality along with the condition that $\frac{Q}{2} \geq q > 0$ implies that $e_1(m) \leq 0$. Similarly, for feature 2, the apportioning of total training error is given by:

$$\begin{aligned} e_2(m) &= \frac{1}{2} tr\_er(\{2\}, m) + \frac{1}{2}(tr\_er(\{1,2\}, m) - tr\_er(\{1\}, m)) \\ &= \frac{1}{2}(Q + q' - q) \geq 0 \end{aligned}$$

The last inequality holds because $Q \geq \frac{Q}{2} \geq q > 0$ and $q' \geq 0$ by construction from Section 2.2. □

Appendix B.

An alternative definition of Shapley value [7] in terms of all possible orders of the players N has been used in the approximation algorithm. Suppose $\pi : \{1, \ldots, n\} \mapsto \{1, \ldots, n\}$ be a permutation and $PermSet(N)$ be the set of all possible permutations with player set N. Given a permutation π, let us denote by $Pred^j(\pi)$ the set of all predecessors of player j in the permutation π, i.e., $Pred^j(\pi) = \{\pi(1), \ldots, \pi(k-1)\}$, if $j = \pi(k)$. Therefore, the Shapley value can be expressed as follows: $\forall j \in N$

$$\phi_j(m) = \sum_{\pi \in PermSet(N)} \frac{1}{n!} \left[v(Pred^j(\pi) \cup \{j\}, m) - v(Pred^j(\pi), m) \right].$$

Synthetic dataset B (sdB): We first generate a binary class label Y from Bernoulli distribution with parameter $p = 0.65$ and then, a 6-dimensional feature vector X for the label Y by drawing a sample such that $X|Y = 1 \sim N([2, 0.4, 2.15, 1, 1.1, 2.05], \Sigma)$ and $X|Y = -1 \sim N([-2, 0.4, -2.15, 1, 1.1, -2.05], \Sigma)$. The matrix Σ is symmetric and most of the entries are zero. The only non-zero entries are $\Sigma_{1,3} = 0.9, \Sigma_{1,6} = 2.6, \Sigma_{3,6} = 2$, $\Sigma_{5,5} = 0.002, \Sigma_{k,k} = 5$, if $k = 1, 3, 6$ and $\Sigma_{l,l} = 0.001$, if $l = 2, 4$. The class conditional means of feature 2,4,5 are the same, and the only differentiating features between two classes are features 1,3, and 6. This is further supported by large variance for the later classes. Thus, features 1,3 and 6 are intuitively important and dominant for label prediction. This is reflected in the SVEA values as for features 1,3, and 6, the 95% intervals on SVEA values lie below zero that is shown in the top left panel of Figure 3.

Algorithm for Shapley value approximation

Require: Feature set $N = \{1, 2, \ldots, n\}$, Number of sample permutations $samPerm$, Number of examples m, Set of coalitions $Sam\_co\_set = [()]$.
Ensure: Approximate Shapley value $\hat{Sh}_j(m) \forall j \in N$
Initialize: $v((), m) = 0$, Shapley value estimate $\hat{Sh}_j(m) := 0 \ \ \forall j \in N$.
Define $tr\_er(\cdot, m)$ on Sam_co_set and compute $tr\_er((), m) = \tilde{c}(m)$ using LP in Section 1.
for $s = 1, 2, .., samPerm$ **do**
  Take $\pi \in PermSet(N)$ with probability $\frac{1}{n!}$.
  for $j = 1, 2, ..., n$ **do**
    Compute the sets $Pred^j(\pi)$ and $Pred^j(\pi) \cup \{j\}$,
    if $Pred^j(\pi)$ **not in** Sam_co_set **then**
      Compute $tr\_er(Pred^j(\pi), m)$.
      Compute $v(Pred^j(\pi), m) = \tilde{c}(m) - tr\_er(Pred^j(\pi), m)$.
      Append $Pred^j(\pi)$ to Sam_co_set.
    end if
    if $Pred^j(\pi) \cup \{j\}$ **not in** Sam_co_set **then**
      Compute $tr\_er(Pred^j(\pi) \cup \{j\}, m)$.
      Compute $v(Pred^j(\pi) \cup \{j\}, m) = \tilde{c}(m) - tr\_er(Pred^j(\pi) \cup \{j\}, m)$.
      Append $Pred^j(\pi) \cup \{j\}$ to Sam_co_set.
    end if
    $\hat{Sh}_j(m) = \hat{Sh}_j(m) + v(Pred^j(\pi) \cup \{j\}, m) - v(Pred^j(\pi), m)$.
  end for
end for
$\hat{Sh}_j(m) = \frac{\hat{Sh}_j(m)}{samPerm}, \ \ \forall j \in N.$

References

[1] M. G. Fiestras-Janeiro, I. García-Jurado, and M. A. Mosquera, "Cooperative games and cost allocation problems," *Top*, vol. 19, no. 1, pp. 1–22, 2011.

[2] S. Khare, B. Khan, and G. Agnihotri, "A Shapley value approach for transmission usage cost allocation under contingent restructured market," in *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*. IEEE, 2015, pp. 170–173.

[3] A. Kimms and I. Kozeletskyi, "Shapley value-based cost allocation in the cooperative traveling salesman problem under rolling horizon planning," *EURO Journal on Transportation and Logistics*, vol. 5, no. 4, pp. 371–392, Dec 2016. [Online]. Available: https://doi.org/10.1007/s13676-015-0087-3

[4] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic game theory*. Cambridge university press, 2007.

[5] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[6] M. J. Osborne and A. Rubinstein, *A course in game theory*. MIT press, 1994.

[7] Y. Narahari, *Game Theory and Mechanism Design (IISc Lecture Notes Series)*. World Scientific Publishing Company / IISc Press, Mar. 2014.

[8] H. P. Young, "Monotonic solutions of cooperative games," *International Journal of Game Theory*, vol. 14, no. 2, pp. 65–72, 1985.

[9] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," *arXiv preprint arXiv:2002.11097*, 2020.

[10] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the Shapley value based on sampling," *Computers & Operations Research*, vol. 36, no. 5, pp. 1726–1730, 2009.

[11] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[12] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[13] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with relieff," *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, 1997.

[14] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 1, no. 25, pp. 1–14, 2013.

[15] B. Cancela, V. Bolón-Canedo, A. Alonso-Betanzos, and J. Gama, "A scalable saliency-based feature selection method with instance level information," *arXiv preprint arXiv:1904.13127*, 2019.

[16] S. Cohen, G. Dror, and E. Ruppin, "Feature selection via coalitional game theory," *Neural Computation*, vol. 19, no. 7, pp. 1939–1961, 2007.

[17] X. Sun, Y. Liu, J. Li, J. Zhu, X. Liu, and H. Chen, "Using cooperative game theory to optimize the feature selection problem," *Neurocomputing*, vol. 97, pp. 86–93, 2012.

[18] X. Sun, Y. Liu, J. Li, J. Zhu, H. Chen, and X. Liu, "Feature evaluation and selection with cooperative game theory," *Pattern recognition*, vol. 45, no. 8, pp. 2992–3002, 2012.

[19] F. Afghah, A. Razi, S. R. Soroushmehr, S. Molaei, H. Ghanbari, and K. Najarian, "A game theoretic predictive modeling approach to reduction of false alarm," in *ICSH*. Springer, 2015, pp. 118–130.

[20] F. Afghah, A. Razi, and K. Najarian, "A shapley value solution to game theoretic-based feature reduction in false alarm detection," *arXiv preprint arXiv:1512.01680*, 2015.

[21] A. Razi, F. Afghah, A. Belle, K. Ward, and K. Najarian, "Blood loss severity prediction using game theoretic based feature selection," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2014, pp. 776–780.

[22] A. Razi, F. Afghah, and V. Varadan, "Identifying gene subnetworks associated with clinical outcome in ovarian cancer using network based coalition game," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 6509–6513.

[23] M. Cavallo and Ç. Demiralp, "A visual interaction framework for dimensionality reduction based data exploration," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[24] S. Liu, J. Yao, C. Zhou, and M. Motani, "Suri: Feature selection based on unique relevant information for health data," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 687–692.

[25] S. Yoon, Y. Song, K. C. Bureau, M. Kim, F. C. Park, and Y.-K. Noh, "Interpretable feature selection using local information for credit assessment," 2018, nIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy.

[26] S. Muñoz-Romero, A. Gorostiaga, C. Soguero-Ruiz, I. Mora-Jiménez, and J. L. Rojo-Álvarez, "Informative variable identifier: Expanding interpretability in feature selection," *Pattern Recognition*, vol. 98, 2020.

[27] K. Leino, M. Fredrikson, E. Black, S. Sen, and A. Datta, "Feature-wise bias amplification," in *International Conference on Learning Representations (ICLR)*, 2019.

[28] E. Strumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, 2014. [Online]. Available: https://doi.org/10.1007/s10115-013-0679-x

[29] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 2016, pp. 598–617.

[30] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[31] A. Ghorbani and J. Zou, "Data Shapley: Equitable valuation of data for machine learning," in *International Conference on Machine Learning*, 2019, pp. 2242–2251.

[32] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos, "Towards efficient data valuation based on the Shapley value," in *AISTATS*, 2019, pp. 1167–1176.

[33] M. Sundararajan and A. Najmi, "The many Shapley values for model explanation," *arXiv preprint arXiv:1908.08474*, 2019.

[34] L. Merrick and A. Taly, "The explanation game: Explaining machine learning models using shapley values," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2020, pp. 17–38.

[35] B. Peleg and P. Sudhölter, *Introduction to the theory of cooperative games*. Springer Science & Business Media, 2007, vol. 34.

[36] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. The MIT Press, 2012.

[37] I. Steinwart and A. Christmann, *Support vector machines*. Springer Science & Business Media, 2008.

[38] U. Faigle and W. Kern, "The Shapley value for cooperative games under precedence constraints," *International Journal of Game Theory*, vol. 21, no. 3, pp. 249–266, 1992.

[39] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms." *Journal of Machine Learning Research*, vol. 18, pp. 174–1, 2017.

[40] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework." *Journal of Multiple-Valued Logic & Soft Computing*, vol. 17, 2011.

[41] D. Dua and C. Graff, "UCI ML repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[43] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *arXiv preprint arXiv:1601.07996*, 2016.

[44] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[45] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[46] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "OpenML: Networked science in machine learning," *SIGKDD Explorations*, vol. 15, no. 2, pp. 49–60, 2013.