IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 April 2024, accepted 2 June 2024, date of publication 11 June 2024, date of current version 9 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3412851

RESEARCH ARTICLE

Efficient Feature Ranking and Selection Using Statistical Moments

YAEL HOCHMA, YUVAL FELENDLER, AND MARK LAST

Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Be'er-Sheva 84105, Israel

Corresponding author: Yael Hochma (Yaelhoc@post.bgu.ac.il)

ABSTRACT Unsupervised feature selection methods can be more efficient than supervised methods, which rely on the expensive and time-consuming data labeling process. The paper introduced skewness as a novel, unsupervised, and computationally efficient feature ranking metric, suitable for both classification and regression tasks. Its feature selection effectiveness is compared to several state-of-the-art supervised and unsupervised feature ranking and selection methods. Both theoretical analysis and empirical evaluation on several popular classification and regression algorithms show that statistical moment-based feature selection algorithms are competitive in terms of accuracy and mean squared error (MSE) with the state-of-the-art supervised approaches for feature ranking and selection, including Fast Correlation Based Filter (FCBF), Minimum Redundancy Maximum Relevance (MRMR), and Mutual Information Maximization (MIM). We also present a mathematical proof based on some common assumptions, which explains the high effectiveness of statistical moments in the feature ranking procedure. Moreover, statistical moment-based feature selection is shown empirically to run faster, on average, than the supervised approaches and the unsupervised Laplacian Score method. Additionally, skewness-based feature selection, in contrast to variance-based selection, does not depend on data normalization that requires additional computational time and may affect the feature ranking results.

INDEX TERMS Feature ranking, unsupervised feature selection, skewness, variance.

I. INTRODUCTION

The tremendous growth in the volume of available data creates challenges for machine-learning algorithms in terms of scalability, processing times, predictive performance, and explainability. Therefore, feature selection is an essential pre-processing step in machine learning tasks when dealing with high-dimensional data. In [1], a risk curve was proposed to qualitatively describe the out-of-sample prediction accuracy of variably parameterized machine learning models. The risk peaks when the number of features becomes close to the sample size. This behavior resembles some key patterns observed in large models and reinforces the fact that it is still important to apply feature selection as a pre-processing step.

Feature selection methods can be classified as supervised [2], semi-supervised [3] or unsupervised [4], depending on the information they use in the feature selection process. Supervised feature selection relies only on labeled data instances, while semi-supervised methods utilize both labeled and unlabeled records. In contrast, unsupervised feature selection (UFS) methods do not require a labeled dataset at all.

Many feature selection algorithms, a.k.a variable selection algorithms, build upon feature ranking as a principle or secondary selection process, since it is scalable, simple, and empirically successful in most cases. In feature ranking algorithms, each one of the original features of a dataset is scored based on some statistical or information-theoretic measures of its importance. Then, the features are sorted by their scores and the top-ranking features can be selected using a predefined threshold. Moreover, in practical situations, the aim is not restricted to predicting the true class of a given observation, but rather it also involves recognizing the input characteristics that play a vital role in a specific behavior. For instance, in the case of a gene expression profile, the initial goal could be to predict the patient's response to a particular

The associate editor coordinating the review of this manuscript and approving it for publication was Xiong Luo.

therapy, while the secondary objective could be to determine the section of the genome that is accountable for the favorable or detrimental reaction [5].

Supervised feature selection methods depend on the availability of true and reliable class labels. However, there are many real-world scenarios where data labeling may be delayed. Moreover, feature selection approaches based on actual labels can be time-consuming and expensive. Consequently, unsupervised feature selection methods are becoming increasingly preferable.

In this work, we explore the use of two statistical moments, variance and skewness, as unsupervised and scalable feature ranking and selection metrics, which are shown theoretically and empirically to produce effective feature subsets.

The main contributions of this paper are:

- We introduce skewness as a novel unsupervised and computationally efficient feature ranking metric, suitable for both classification and regression tasks.
- We compare the effectiveness of variance-based and skewness-based feature selection procedures to supervised FS methods using 40 benchmark datasets, four classification algorithms, and three regression algorithms.
- We present a mathematical proof that under some common assumptions, the probability distribution of a predictive variable is related to its discriminative power, which may explain the high effectiveness of statistical moments in the feature ranking procedure.
- Finally, we discuss the advantages of using skewness rather than variance for unsupervised feature ranking.

This paper is organized as follows: First, we present a literature review, then we present the goal of this work and provide a formal problem definition. Next, we describe the statistical moment-based feature ranking and selection methodology and explore its effectiveness using both mathematical analysis and evaluation experiments. Finally, we present our main conclusions and suggested directions for future research.

II. RELATED WORK

We here go through relevant topics in the scope of feature selection and more specifically, feature ranking as a primary or a secondary feature selection mechanism. The first sub-section explains the motivation of applying such methods and reviews some feature selection approaches for classification and regression tasks, using labeled and unlabeled data. Since the proposed feature selection method is based on feature ranking, sub-section II-C discusses feature ranking methods focusing on filter-based approaches and their use.

A. FEATURE SELECTION OVERVIEW

Feature selection is a process of choosing a subset of original features such that the feature space is optimally reduced according to certain evaluation criteria [6].

We are facing a continuous increase in data volumes, both in terms of the number of instances and number of features in such applications as genome projects [7], text categorization [8], image retrieval [9], and customer relationship management [10]. This data abundance may cause some problems for scalability and learning performance of many machine-learning algorithms, especially when high-dimensional data contains a significant amount of irrelevant and redundant features. Therefore, feature selection is crucial for machine learning tasks handling high-dimensional data.

However, there are claims that the use of feature selection methods in the big data domain is unnecessary due to the superior performance of deep neural networks, which usually consider all original features. Still, the benefits of feature selection include better explainability of the induced machine learning models along with reducing the data collection costs as well as decreasing the training and the inference times.

Feature selection algorithms fall into three broad categories: filter, wrapper, and embedded approaches [11]. Filter methods assign a score to each feature to indicate its importance, such as entropy, information gain, chi-square test etc. [12], [13]. Several methods have been proposed for discovering the relations between the input variables and the output, while the most familiar and common are mutual information-based approaches [14]. These approaches, have an incremental nature, also called greedy, which means they are prone to sub-optimal decisions. One of the first feature selection methods based on information theory is the Mutual Information Maximization (MIM) [15], which adopts mutual information to measure the association between each feature and the output class vector, and do not consider the interaction between features. One of the most successful and well-known MI-based approach is the Minimum Redundancy Maximum Relevance (MRMR) framework [16], which finds at each step the feature with the maximum relevance to the target class and the minimum redundancy with the previously selected features. Torkkola [17] suggested another filter method for constructing features using a mutual information criterion. The author maximizes I(φ,y) for m dimensional feature vectors φ and target vectors y. Fast Correlation-Based Filter (FCBF) is an example of a filter-based feature selection approach which exploits feature-class correlation and feature-feature correlation simultaneously. The algorithm selects a subset of features that are highly correlated with the class labels and removes redundant features by calculating the symmetric uncertainty of the input features and the class label [6]. Wrapper-based approach for feature selection requires one predetermined learning algorithm and uses its performance to evaluate and determine the features subset. Some examples are recursive feature elimination and genetic algorithms [18]. In embedded techniques, the feature selection algorithm is integrated as part of the learning algorithm; the most common embedded technique is LASSO algorithm [19].

For each new subset of features, the wrapper-based approach learns a hypothesis (or a classifier) so it tends to find features better suited to the predetermined learning algorithm, resulting in superior learning performance of that algorithm. Unfortunately, it also tends to be more computationally expensive than the filter-based models [6]. Once the number of features becomes very large, filter-based methods are usually preferred due to their lower computational complexity.

B. UNSUPERVISED FEATURE SELECTION

According to the information utilized by the algorithms we can classify feature selection methods as supervised [2]; [19] and unsupervised [4]. The first category requires a set of labeled data, whereas Unsupervised Feature Selection (UFS) methods [4], [20] do not require a labeled dataset.

Due to the limited amounts and high costs of labeled data, UFS methods have attracted significant interest in the machine learning community. The UFS methods have two important advantages. (1) They are unbiased and perform well when prior knowledge is not available. (2) They can reduce the risk of data overfitting in contrast to supervised feature selection methods, which may be unable to deal with a new class of data [21]. Max variance is one method for unsupervised filter-based feature selection. This method selects a subset of features based on a user-specified threshold, e.g., keeps the top k features with the largest variance. It assumes that features with higher variance are more useful for classification and regression tasks. Xu et al. [22] propose an unsupervised filter-based gene selection framework by applying diffusion maps to address the multi- dimensionality problem and using the eigenfunctions of Markov matrices as a coordinate system on the original dataset in order to obtain efficient representation of data geometric descriptions. The authors applied three types of gene selection before applying diffusions: correlation coefficient of a single variable to other variables, max variance and selection of variables with a bimodal (double hump) probability density. They found that applying these standard, filter–based feature selection methods achieve success. SPECtral feature selection (SPEC) [21] ranks each feature by three different metrics through spectral analysis, Hou et al. propose a general framework for feature selection termed as Joint Embedding Learning and Sparse Regression (JELSR) [23]. In Laplacian Score [24], the importance of a feature is evaluated by its variance and its power to preserve locality. This method assigns high weights to features that can best preserve the underlying manifold structure represented by the Laplacian matrix. Laplacian score uses a nearest-neighbor graph to model the local geometric structure of the data. This idea is based on the assumption that observations that are close to each other are probably related to the same cluster. Thus, those features that have similar values for close objects and distant values for remote ones are the most relevant features. Another widespread choice is the pseudo-label-based methods. These methods usually generate pseudo labels from data through clustering algorithms and then, select features based on their utility in predicting the pseudo labels with sparse learning-based framework. An example of this kind of methods is MCFS [25]. In [26], the authors rank features using auto-encoders and evaluate the overall reconstruction error of the auto encoder in absence of any specific feature. Their assumption is that a low error indicates that a specific feature is unimportant for representing the sample, or may be highly correlated with other present features.

In unsupervised case, max variance is the most practical method for working with big data, which has a linear time complexity in terms of the number of features and the number of instances.

C. FEATURE RANKING

In the feature ranking-based approach, each feature of a dataset is scored based on one or several statistical or information-theoretic measures. Then, the features are ranked based on their score and the top ranking features are selected as the predictive features using a predefined threshold that determines the number of features to be selected from a dataset.

Examples of feature ranking-based methods include Chi-Square-Based Feature Selection (CQFS) [27] as well as information-theoretic measures such as information gain, gain ratio, Pearson correlation etc. As mentioned in [3] it is still an open problem to determine the optimal number of selected features. In practice, one usually adopts a heuristic way to search through the size of the subset of features and choose the number that provides the best classification performance.

It is noteworthy that feature ranking-based methods take less runtime but fail to remove redundant features [28]. To address this limitation one can use a suitable redundancy analysis approach as well.

In [29], the proposed Fast Hybrid Feature Selection based on Correlation-Guided Clustering (FHFS-CGC) method combines correlation-guided clustering and particle swarm optimization to select the most relevant features from high-dimensional datasets. This fully supervised method first clusters the features based on pairwise correlations and then applies particle swarm optimization to select the most discriminative feature from each cluster. The computational complexity of their algorithm is O $[(D-1)!/(2^{D-2})]$, where D is the number of features.

In [30], the authors proposed two approaches for semi-supervised learning of feature rankings with several classification and structured output prediction tasks including multi-label classification, hierarchical multi-label classification, and multi-target regression. The proposed methods are based on predictive clustering tree ensembles and the Relief family of feature ranking algorithms and they are evaluated on static datasets only. Semi-supervised learning methods make use of unlabeled, in addition to labeled data under the assumption that clusters of unlabeled samples resemble the distribution of class labels. When this assumption holds, ignoring unlabeled data may be counter-productive. In contrast, we explore here a fully unsupervised approach to feature ranking, which does not rely on the class labels of any samples. We evaluate a new feature ranking metric, the third statistical moment of feature values, and compare its effectiveness to several feature impurity functions including feature variance used by [30].

III. PROBLEM DEFINITION

The goal of this work is to develop and evaluate unsupervised and computationally efficient filter-based approaches to feature ranking and selection, which can be used to induce reasonably accurate classification and regression models.

Problem 1: (Unsupervised Feature Ranking and Selection): Given a dataset D, a feature scoring function S, and the number of features to be selected k, the problem of unsupervised feature selection aims at selecting a subset of k most informative features from the original feature space n ($k < n$) *without making use of instance labels. In particular, consider a set of d examples,* x_i, y_i ($i = 1 \ldots d$) *consisting of n numeric features,* $x(i,j)$ ($j = 1, \ldots n$)*. Unsupervised feature selection via variable ranking is using a scoring function* $S(j)$ *computed only from the values* $x(i,j)$*. After sorting the variables in descending order of their* $S(j)$ *scores, the top k variables can be used as predictive features by a classification or regression algorithm.*

IV. METHODOLOGY

We are proposing skewness as a novel unsupervised feature scoring function and then proceed with evaluating both variance-based and skewness-based feature selection approaches.

Definition 1: Skewness (third statistical moment) is a measure of symmetry, or more precisely, the lack of symmetry (known as asymmetry). Data distribution is symmetric if it looks the same to the left and right of the center point [31]. Positive skewness indicates that the mean is higher than the median, whereas, in a distribution that is negatively skewed, the mean is lower than the median. In cases of a symmetric data histogram, such as the normal distribution, a feature has zero skewness. As indicated by [32], there are asymmetric features where only the minority of an attribute's values strongly point to one of the target classes. This empirical phenomenon may be explained by the normal distribution of most "weak" (noisy) features [1], which are affected by multiple independent hidden factors, as opposed to "strong" (predictive) features, which are characterized by skewed distributions. The skewness of a random variable X, or the third standardized moment γ_1, is defined mathematically as:

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \tag{1}$$

μ is the central moment and σ is the standard deviation

It is worth mentioning that skewness values remain stable across scaling or normalization of datasets.

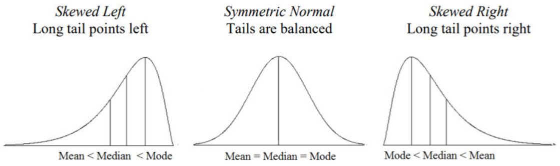


FIGURE 1. Sketches showing general position of mean, median, and mode in a population [33].

Definition 2 (Mean): The sample mean is the average of the values of a variable in a sample, which is the sum of those values divided by the number of values. Using mathematical notation, if a sample of N observations on variable X is taken from the population, the sample mean is:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

The skewness values vary between plus and minus infinity. In our feature ranking and selection experiments, we evaluated feature scoring both by the actual and the absolute skewness values though features having positive skewness values are much more prevalent than the negatively skewed ones. Examples of positively skewed features include personal incomes, waiting times in a queue, or the life span of a technical device, in contrast to highly skewed negative features like student scores in an easy exam, where there are very few failures.

Definition 3: variance (second statistical moment) of a random variable is the expectation of the squared deviation of variable values from its mean. In practice, it measures how far the feature values are spread out from their average value. It is usually assumed that features with higher variance may contain more information needed to discriminate between class labels [24] though we are not aware of any formal proof of this common assumption.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} \tag{3}$$

μ is the is the average value and n is the number of samples.

To use variance for feature selection, we should normalize the features to the same scale as the raw values of variance are sensitive to scaling unlike skewness-based scores, which do not require normalization before applying them to features of different scales.

V. THE IMPACT OF PREDICTIVE VARIABLES DISTRIBUTION ON FEATURE RANKING

This section analyzes how the distribution of a predictive variable may affect its discriminative power in binary classification problems, which fit the popular Logistic Regression model. The logistic regression model assumes a log-linear relationship between the predictor variables and the binary outcome. Mathematically, this relationship can be expressed in terms of the logistic function, which transforms a linear predictor (i.e., a weighted sum of the predictor variables) into

Algorithm 1 Statistical Ranking-Based Feature Selection [34]

```
Input: $k \Leftarrow$ number of top features to be selected and $X \Leftarrow$ Dataset and $M \Leftarrow$ statistical feature scoring function (e.g., variance or skewness)
Output: Subset of $k$ selected features
  1) for each feature $x$ in $X$
       2) Calculate feature score M(x)
  3) return Array of -feature ID, score
  4) TotalRank $\Leftarrow$ Sort feature scores in descending order
  5) return top $k$ features from TotalRank
```

a probability value between 0 and 1. The logistic function takes the form:

$$m(x) = \frac{\exp^{\beta x}}{1 + \exp^{\beta x}} \tag{4}$$

where $m(x)$ is the probability of a binary outcome (usually denoted as 0 or 1), x is the value of a predictor variable, β is the coefficient (i.e., the weight) associated with the predictor variable, and exp represents the exponential function. Accordingly, the complement of the above equation will represent the probability of the alternative outcome (either 1 or 0, respectively) as follows:

$$1 - m(x) = \frac{1}{1 + \exp^{\beta x}} \tag{5}$$

The logistic regression model builds upon several key assumptions, including linearity, independence, and homogeneity of variance. Linearity refers to the assumption that the relationship between the predictor variable values and the binary outcome is log-linear, as captured by the logistic function. Independence refers to the assumption that the observations (i.e., the instances) are independent of each other, meaning that the probability of an event occurring for one instance does not depend on the probabilities of events occurring for other instances. Homogeneity of variance refers to the assumption that the variance of the probability distribution of the binary outcome is constant across all predictor variables.

If the two classes are linearly separable given the values of a predictor variable x, there is a threshold value Th such that all instances in the left interval $x \leq Th$ are labeled by 0, whereas in the right interval $x > Th$ all instances are labeled by 1 (or vice versa). Such a variable should be assigned the highest score by a supervised impurity metric, such as the Information Gain or the Twoing, as a 'perfect predictor' of the binary class. In contrast, the variables having the same proportion of each class label in both intervals defined by any threshold value should have the lowest rank as irrelevant features having 'zero impurity'. In Theorem 1 below, we prove that if a numeric feature fits the logistic regression model, its variance and skewness, which can be estimated without knowing the instance labels, can indicate its actual discriminative power, justifying the use of variance and skewness as an effective feature ranking metric in unsupervised scenarios.

Theorem 1: If a binary classification function of a numeric predictive variable fits the logistic regression model, the likelihood of having the same class proportions in both intervals resulting from any variable split will be inversely proportional to the feature variance and skewness.

Proof: The expected number of instances belonging to each binary outcome in a given interval can be calculated by integrating the logistic function over the interval range. Let Th be a threshold value between the two intervals. The expected number of instances belonging to the binary outcome of 1 in the lower interval ranging from the minimum feature value *min* to Th is:

$$\text{Expected}(N_1/x \leq Th) = \int_{\text{min}}^{\text{Th}} \frac{\exp(\beta x)}{1 + \exp(\beta x)}, dx \tag{6}$$

Accordingly, the expected number of instances belonging to the same outcome of 1 in the upper interval ranging from Th to *max* is:

$$\text{Expected}(N_1/x > Th) = \int_{\text{Th}}^{\text{max}} \frac{\exp(\beta x)}{1 + \exp(\beta x)}, dx \tag{7}$$

The difference between the expected proportions of instances belonging to the binary outcome of 1 in the two intervals can be calculated as:

$$\frac{Expected(N_1/x > Th)}{\int_{\text{Th}}^{\text{max}} dx} - \frac{Expected(N_1/x \leq Th)}{\int_{\text{min}}^{\text{Th}} dx} \tag{8}$$

Assuming that the feature range [*min*; *max*] is fixed following the min-max normalization and given any threshold value Th, the above difference between the expected proportions will increase, thus reducing the likelihood of zero impurity, if most feature values in the lower interval will approach *min*, whereas most feature values in the upper interval will approach *max*. Such a two-sided "fat tail" distribution will result in a higher feature variance, whereas a one-sided "fat tail" would also result in a higher absolute skewness. The feature variance is calculated by:

$$\int_{\text{min}}^{\text{max}} x^2 dx - (\int_{\text{min}}^{\text{max}} x dx)^2 \tag{9}$$

This completes the proof.

Though in real-world datasets predictive features do not necessarily fit the logistic regression model, Theorem 1 explains why unsupervised feature ranking based on statistical moments, such as variance and skewness, may be nearly as effective as the supervised approaches. In sub-section VII-C1, we explore the effect of feature distribution on supervised and unsupervised metrics using sample features from the Sonar dataset.

VI. COMPUTATIONAL TIME COMPLEXITY

Based on the methodology presented above, we present here the computational analysis of the Statistical Ranking-based Feature Selection algorithm vs. four benchmark feature selection algorithms (Fast Correlation Based Filter (FCBF)

TABLE 1. Summery of computational time complexity for the proposed and benchmark feature selection algorithms.

Ranking Method	Time Complexity
Skewness	$O(DN)$
Variance	$O(DN)$
FBCB	$O(NDlog(D))$ [6]
MRMR	$O(DN^2)$ [16]
MIM	$O(kDCN)$ [37]
Laplacian score	$O(DN^2)$ [24]

Minimum Redundancy Maximum Relevance (MRMR), Mutual Information Maximization (MIM) and Laplacian Score(LS).

Feature ranking by skewness and variance (see algorithm 1) metrics calculation has a linear time complexity in terms of the number of features D because we iterate over all original features. For each feature, skewness and variance metrics calculation is linear in terms of the number of instances N in a dataset [35], [36]. Therefore, the overall complexity of the moment-based FS is only $O(DN)$, which, as shown below, is lower than the complexity of state-of-the-art supervised FS algorithms. It is worth mentioning that normalization of the data is required before ranking features based on variance; the normalization step increases the computational effort, which is saved with skewness-based ranking. Thus, according to [6], the supervised Fast Correlation Based Filter algorithm has a time complexity of $O(NDlogD)$. In another supervised algorithm, MRMR, the Mutual Information of all possible feature pairs: feature-feature and feature-class, is computed. Therefore, the computational complexity is quadratic $O(DN^2)$, since the number of distinct classes is bounded by the number of instances. The time complexity of calculating mutual information,is $O(CN)$ in terms of the number of instances N and number of classes C, because all instances need to be examined for probability estimation. Since MIM calculates only the MI between feature-class pairs, its time complexity can be calculated by $O(kDCN)$ [37]. The time complexity of Laplacian Score is dominated by the cost of building a nearest neighbor graph which is quadratic in the size of the training set. Hence, its time complexity is $O(DN^2)$

VII. EXPERIMENTS

In this section, we perform experiments on a variety of benchmark datasets to evaluate the effectiveness of statistical moment-based feature selection. We describe the datasets and the experimental settings before presenting the details of the experimental results.

A. DATASETS

To evaluate and compare the performance of the proposed feature selection algorithm, we conducted an extensive experimental study using a total of 40 benchmark datasets. These datasets were downloaded from the widely-used UCI repository [38] and were carefully selected to contain a diverse range of data characteristics and complexities. Specifically, we used 30 UCI datasets suitable for the classification task and 10 datasets suitable for the regression task. On some of the datasets, we conducted simple preprocessing: missing values were replaced by the mode for categorical features and by the mean for continuous features. Statistical moments were calculated for categorical features by converting them into numerical labels using the label encoder [39].

We present our findings based on the classification and regression datasets separately. Our experiments with both types of datasets demonstrate the applicability of the proposed feature ranking techniques to a broad range of machine-learning tasks.

1) CLASSIFICATION DATASETS DESCRIPTION

The statistics of these datasets are summarized in Table 2. The datasets contain between 101 to 45211 instances and between 6 to 19,993 features, both numerical and categorical, mostly balanced.

2) DATASETS DESCRIPTION FOR REGRESSION TASKS

We used 10 benchmark datasets suitable for regression tasks and available from the UCI repository [38]. The statistics of these datasets are summarized in Table 3. The datasets contain between 204 to 241,600 instances and between 9 to 128 features, both numerical and categorical.

B. EXPERIMENTAL SETTINGS FOR FEATURE SELECTION METHODS

We compare our moment-based framework with the following unsupervised and supervised feature selection methods, using the scikit-feature package implementation [40], with their default settings:

- Laplacian Score (LS) [24] is a state-of-the-art unsupervised feature selection method that selects features that can best preserve the local manifold structure of the data. The method constructs the affinity matrix W, which represents the similarity between each pair of samples. In our experiments, it is constructed using the K-nearest Neighbor (KNN) graph method, as the default setting, where the K nearest neighbors are identified based on the Euclidean distance between samples, and an edge is created between each pair of samples in the graph. The Laplacian Scores are then calculated for each feature using W, and the features are ranked based on their corresponding Laplacian Scores.
- Fast Correlation Based Filter (FCBF) [6]: a state-of-the-art supervised feature selection method, which selects features that are highly correlated with the class labels and removes redundant features by calculating the symmetric uncertainty of the input features and the class label. The method has a hyperparameter called Delta (float) - a threshold parameter with a default value of 0.

TABLE 2. **Datasets description for classification tasks.**

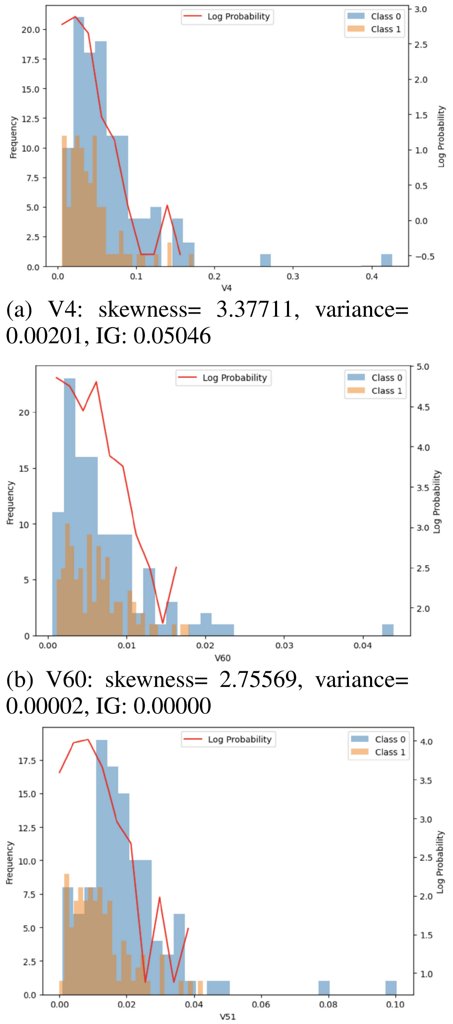
Dataset	Features	Continuous	Categorical	Instances	Classes	majority class
Arrhythmia	451	279	120	159	2	0.54
Bank	45211	14	0	14	2	0.88
Biodeg	1055	42	42	0	2	0.55
Breast-cancer-wisconsin	286	9	0	9	2	0.60
Car	1727	6	0	6	4	0.69
Chess	3195	36	0	36	2	0.52
CNAE-9	1080	856	855	1	9	0.11
CTGs	2126	22	22	0	3	0.60
Darwin	174	450	423	27	2	0.51
Diabetes	768	8	8	0	2	0.65
Digits	10991	16	0	16	10	0.11
Dis	2799	28	0	28	5	0.58
Dry_Bean	13611	16	16	0	7	0.26
EEG Eye State	14980	14	14	0	2	0.55
Glass	214	9	9	0	7	0.30
Heart stalog	270	13	6	7	2	0.54
Heart-c	303	13	4	9	2	0.54
Hepatitis	155	18	1	17	2	0.60
Ionosphere	351	34	34	0	2	0.64
Lymphography	148	18	0	18	4	0.50
Musk	6598	166	166	0	2	0.60
Page-blocks	5473	10	4	6	5	0.89
Phishing dataset	11054	30	0	30	2	0.55
SMK-CAN-187	187	19993	19993	0	2	0.51
Sonar	208	60	60	0	2	0.53
Spambase	4601	57	57	0	2	0.55
Thyroid	9171	28	0	28	2	0.88
Titanic	891	9	6	3	2	0.88
Wine	178	13	13	0	3	0.39
Zoo	101	16	0	16	7	0.20

TABLE 3. **Regression datasets description.**

Dataset	Features	Continuous	Categorical	Instances
Automobile	204	28	15	13
Battery_RUL	15064	9	9	0
Cargo	3943	98	26	72
CO2 Emissions Canada	7385	12	5	7
Communities	1993	128	128	0
Forest Fires	517	13	11	2
House Price Multifeatures	81747	96	61	35
Sgemm Product	241600	18	18	0
Song Poplarity	18835	15	14	1
Super Conductivity	21263	82	82	0

- MRMR [16]: a state-of-the-art supervised feature selection method, uses mutual information as a filter method in order to obtain maximum classification performance with a minimal subset of variables by reducing the redundancies among the selected variables to a minimum.
- MIM [15] a state-of-the-art supervised feature selection method adopts mutual information to measure the relevancy between each feature and the output class vector.

Since unsupervised moment-based methods use the filter approach, their effectiveness may differ across classifiers. In order to reduce the bias of a specific classifier and test the robustness of the evaluated methods, we measure the classification accuracy of four different and commonly used classifiers: k-Nearest neighbors (KNN), Naïve-Bayes (NB), linear Support Vector Machine (linear-SVM) and MLP neural network (MLP), all with their default parameters using 10-fold cross-validation. We present the average classification accuracy obtained for each classifier over a subset of top k selected features, where k varies from 1 to 30 at increments of 1 (note that, if the number of original features of the dataset is less than 30, then k goes up to the number of original features). This approach facilitated the avoidance of biases associated with the selection of specific values of k. In order to rank features with variance metric we normalize the datasets with Min-Max scalar. The Python 3 machine learning software, scikit-Learn [41], and the scikit-feature library [40] were used for the implementation of the evaluated supervised and unsupervised feature selection



(a) V4: skewness= 3.37711, variance= 0.00201, IG: 0.05046

(b) V60: skewness= 2.75569, variance= 0.00002, IG: 0.00000

(c) V51: skewness= 2.69643, variance= 0.00014, IG: 0.05643

FIGURE 2. **Histograms of features having the highest skewness in the sonar dataset.**

methods and the evaluated classifiers with their default parameters.

C. EXPERIMENTAL RESULTS

1) FEATURE RANKING USING VARIANCE AND SKEWNESS

As an example, we examined the difference between feature ranking using the second and the third statistical moments on the Sonar dataset. While the top three features of the Sonar dataset that demonstrate the highest skewness were $V4$, $V60$, and $V51$, the top three features having the highest variance were $V20$, $V17$, and $V21$. The histograms of these six features are shown in Figures 2 and 3. Each histogram was crafted using 30 bins and the red line represents the log probability for class 1.

Before calculating the Information Gain (IG) of numeric features, discretization was performed using the Fayyad and Irani MDL-based algorithm [42]. The method has been implemented by employing MDLP (Minimum Description Length Principle) discretizer from the $mdlp - discretization$ package. To compute the information gain, we calculate the mutual information between each discretized numeric or nominal feature and the target variable using the *mutual_info_classif* function from Scikit-learn [40].

The histograms of the features with the highest skewness (Figures 2a -2c) are characterized by a long tail composed of a few large values leaning to the right, whereas most other values are relatively close to the mean, resulting in a very low variance. However, two out of the three top-skewness variables have high Information Gain values indicating their good discriminative power. In contrast, the histograms of features with the highest variance (shown in Figures 3a - 3c) exhibit a more symmetric appearance. However, two out of three top-variance features have near-zero Information Gain values indicating their apparent irrelevance for the classification task.

Another characteristic of the top skewness features in Figure 2 is the relative monotonicity of the odds ratio between the two classes as opposed to the top variance features in Figure 3. This observation supports the claim of Theorem 1 that for a binary classification function fitting the logistic regression model, the high skewness of feature values caused by a one-sided ‘‘fat tail’’ is an indicator of good discriminative power.

In contrast, the top variance features shown in Figure 3 do not exhibit monotonicity in the odds ratio between the classes. In two out of three cases, their histograms are not characterized by ‘‘fat tails’’, which explains the near-zero values of their Information Gain despite their relatively high variance.

These sample results confirm the claim of Theorem 1 and suggest that in the absence of class labels, feature skewness could potentially serve as an effective feature selection criterion alongside the traditional variance metric.

2) ACCURACY

Tables 4-7 show the results of the evaluated feature selection methods on 30 benchmark datasets using four different classifiers.

Since, as mentioned above, the skewness values range between plus and minus infinity, we examined the effectiveness of feature ranking by absolute skewness values. However, the results of our experiments have shown that ranking the features by their actual skewness values provides better results in terms of classification accuracy, i.e. positive skewness is more informative for discrimination between class labels than negative skewness. For example, the Wine dataset obtained an accuracy score of 0.618 when using absolute skewness values, compared to 0.743 with the actual values. In our future work, we intend to explore further the apparent advantage of positively skewed predictive features, such as income, over negatively scored features, such as the retirement age.

We compared the accuracy of the unsupervised variance and skewness-based feature selection methods to FCBF, MIM

TABLE 4. Average classification accuracy of SVM on 30 data sets, k = [1,30].

Dataset	LS	MRMR	FCBF	MIM	Skewness	Variance
Arrhythmia	0.546	0.553	**0.566**	0.561	0.543	0.551
Bank	0.881	0.883	0.883	0.883	**0.883**	0.883
Biodeg	0.684	0.674	0.699	**0.699**	0.687	0.674
Breast-cancer-wisconsin	0.917	0.914	0.897	0.921	0.928	**0.943**
Car	0.703	0.703	**0.714**	0.703	0.703	0.713
Chess	0.579	0.610	**0.778**	0.603	0.572	0.676
CNAE-9	0.257	0.230	0.328	0.219	**0.510**	0.250
CTGs	0.636	0.570	**0.883**	0.767	0.798	0.604
Darwin	0.565	0.660	**0.715**	0.651	0.512	0.711
Diabetes	0.637	**0.670**	0.625	0.664	0.638	0.642
Digits	0.652	0.663	**0.834**	0.645	0.795	0.786
Dis	0.585	0.594	**0.657**	0.583	0.598	0.600
Dry Bean	0.813	0.842	0.849	0.874	**0.875**	0.867
EEG_Eye_State	0.504	0.498	0.499	0.505	**0.520**	0.498
Glass	0.450	0.439	0.354	0.443	0.437	**0.471**
Heart_stalog	0.652	0.671	**0.756**	0.620	0.656	0.580
Heart-c	0.683	0.601	**0.716**	0.636	0.656	0.549
Hepatitis	0.790	0.765	0.758	**0.794**	0.788	0.711
Ionosphere	0.668	0.782	0.714	**0.797**	0.660	0.603
Lymphography	0.665	0.594	**0.708**	0.538	0.678	**0.708**
Musk	0.763	**0.763**	0.757	0.748	0.696	0.749
Page-Blocks	0.914	0.911	0.915	0.917	0.909	**0.920**
Phishing_dataset	**0.691**	0.566	0.579	0.566	0.632	0.614
SMK-CAN-187	**0.559**	0.550	0.568	0.530	0.535	0.595
Sonar	0.601	0.571	0.543	0.582	0.599	**0.635**
Spambase	0.622	0.663	0.666	0.665	**0.703**	0.683
Thyroid	0.599	0.599	0.610	0.601	**0.630**	0.604
Titanic	0.725	0.725	0.725	0.725	**0.779**	0.778
Wine	0.833	**0.850**	0.634	0.793	0.743	0.634
Zoo	0.702	**0.816**	0.484	0.812	0.556	0.644

TABLE 5. Average classification accuracy of KNN on 30 data sets, k = [1,30].

Dataset	LS	MRMR	FCBF	MIM	Skewness	Variance
Arrhythmia	0.542	0.529	**0.570**	0.534	0.390	0.522
Bank	**0.945**	0.883	0.893	0.883	0.883	0.867
Biodeg	0.582	0.525	0.628	0.489	**0.647**	0.610
Breast-cancer-wisconsin	0.769	0.911	0.896	**0.931**	0.926	0.920
Car	0.699	0.722	0.754	0.722	**0.764**	0.762
Chess	0.548	0.578	**0.730**	0.582	0.563	0.653
CNAE-9	0.197	0.121	0.128	0.115	**0.221**	0.220
CTGs	0.704	0.740	**0.913**	0.790	0.723	0.724
Darwin	0.661	0.645	0.708	0.694	**0.789**	0.712
Diabetes	0.618	0.691	0.678	**0.696**	0.644	0.653
Digits	0.715	0.717	0.865	0.730	**0.836**	0.825
Dis	0.501	0.516	**0.606**	0.561	0.562	0.565
Dry Bean	0.813	0.838	0.850	**0.868**	0.861	0.862
EEG_Eye_State	0.538	0.531	0.514	0.518	0.522	**0.543**
Glass	0.514	0.477	0.465	0.444	**0.586**	0.565
Heart_stalog	0.634	0.600	**0.709**	0.566	0.631	0.626
Heart-c	**0.692**	0.594	0.667	0.559	0.642	0.624
Hepatitis	0.773	0.785	0.770	**0.789**	0.751	0.740
Ionosphere	0.774	**0.843**	0.839	0.807	0.811	0.822
Lymphography	0.626	0.526	0.655	0.523	0.588	**0.642**
Musk	**0.793**	0.726	0.743	0.711	0.621	0.727
Page_Blocks	0.932	0.935	0.939	0.947	**0.955**	0.945
Phishing_dataset	**0.616**	0.530	0.516	0.530	0.594	0.540
SMK-CAN-187	0.548	0.539	0.548	0.527	0.580	**0.587**
Sonar	0.545	0.579	0.541	**0.589**	0.581	0.565
Spambase	0.449	0.436	0.562	0.449	0.675	**0.693**
Thyroid	0.548	0.537	**0.581**	0.571	0.532	0.574
Titanic	0.687	0.708	0.703	0.696	**0.779**	0.745
Wine	0.832	**0.864**	0.700	0.748	0.658	0.698
Zoo	0.720	**0.820**	0.468	0.820	0.630	0.658

and MRMR, the supervised feature selection algorithms, using Wilcoxon signed-rank test. The results show that the difference between the skewness and each other feature selection method is not statistically significant (alpha = 0.05)

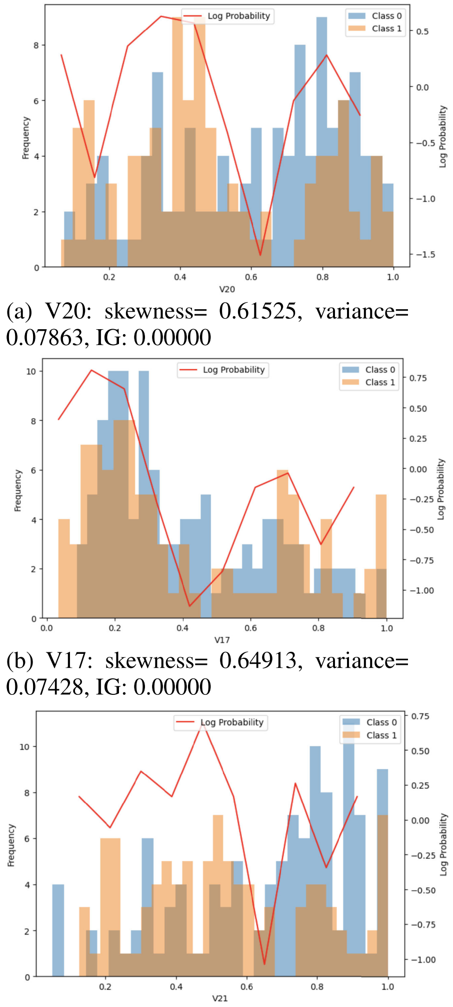
TABLE 6. Average classification accuracy of NB on 30 data sets, k = [1,30].

Dataset	LS	FCBF	MIM	MRMR	Skewness	Variance
Arrhythmia	**0.563**	0.180	0.116	0.257	0.117	0.305
Bank	0.875	0.872	0.861	0.864	**0.881**	0.600
Biodeg	**0.665**	0.492	0.547	0.652	0.618	0.591
Breast-cancer-wisconsin	0.913	0.914	0.898	0.918	0.878	**0.944**
Car	0.636	0.596	0.629	0.629	**0.651**	0.600
Chess	0.535	0.651	0.602	0.605	0.547	**0.674**
CNAE-9	0.156	0.130	0.127	0.115	**0.310**	0.229
CTGs	0.778	0.733	**0.923**	0.867	0.768	0.771
Darwin	0.507	0.697	0.657	0.652	0.667	**0.720**
Diabetes	0.671	**0.745**	0.698	0.756	0.681	0.723
Digits	0.589	**0.711**	0.561	0.595	0.677	0.690
Dis	0.190	0.142	0.221	0.116	0.084	**0.324**
Dry Bean	0.775	0.828	**0.862**	0.820	0.806	0.813
EEG_Eye_State	0.465	0.462	0.472	0.492	**0.552**	0.441
Glass	0.416	0.170	0.254	0.262	**0.479**	0.463
Heart_stalog	0.737	0.723	**0.795**	0.711	0.695	0.682
Heart-c	0.737	0.728	**0.782**	0.711	0.669	0.663
Hepatitis	0.451	0.743	0.752	0.645	**0.753**	0.752
Ionosphere	0.702	**0.782**	0.722	0.763	0.690	0.593
Lymphography	0.474	0.478	0.645	0.419	0.422	**0.687**
Musk	0.775	0.757	0.824	0.733	0.753	**0.827**
Page-Blocks	0.901	0.850	**0.913**	0.883	0.859	0.909
Phishing_dataset	0.678	0.566	0.576	0.566	**0.769**	0.620
SMK-CAN-187	0.607	0.560	0.613	0.586	**0.762**	0.602
Sonar	0.582	0.588	0.568	0.609	**0.660**	0.610
Spambase	0.527	0.648	0.603	0.653	0.661	**0.687**
Thyroid	0.099	0.111	0.055	0.059	0.213	**0.610**
Titanic	0.636	0.620	0.593	0.646	0.754	**0.778**
Wine	0.848	0.843	0.820	0.865	**0.871**	0.798
Zoo	0.688	0.848	0.360	**0.850**	0.628	0.626

TABLE 7. Average classification accuracy of MLP Classifier on 30 data sets, k = [1,30].

Dataset	LS	FCBF	MIM	MRMR	Skewness	Variance
Arrhythmia	0.468	0.499	0.521	**0.561**	0.493	0.451
Bank	0.881	**0.885**	0.883	0.884	0.865	0.857
Biodeg	**0.665**	0.492	0.547	0.652	0.618	0.591
Breast-cancer-wisconsin	0.913	0.914	0.898	0.918	0.878	**0.944**
Car	0.684	0.521	0.576	0.703	**0.723**	0.451
Chess	0.579	0.610	0.778	0.603	**0.865**	0.857
CNAE-9	0.156	0.130	0.127	0.115	**0.310**	0.229
CTGs	0.778	0.733	**0.923**	0.867	0.768	0.771
Darwin	0.565	0.660	0.702	0.651	**0.713**	0.709
Diabetes	0.671	**0.745**	0.698	0.756	0.681	0.723
Digits	0.652	0.673	0.634	0.645	**0.698**	0.674
Dis	0.532	0.574	**0.647**	0.583	0.644	0.587
Dry Bean	**0.878**	0.832	0.821	0.877	0.864	0.864
EEG_Eye_State	0.465	0.462	0.472	0.492	**0.552**	0.441
Glass	0.416	0.170	0.254	0.262	**0.479**	0.463
Heart_stalog	0.737	0.723	**0.795**	0.711	0.695	0.682
Heart-c	0.737	0.728	**0.782**	0.711	0.669	0.663
Hepatitis	0.451	0.743	0.752	0.645	**0.753**	0.752
Ionosphere	0.702	**0.782**	0.722	0.763	0.690	0.593
Lymphography	0.474	0.478	0.645	0.419	0.422	**0.687**
Musk	0.775	0.757	0.824	0.733	0.753	**0.827**
Page-Blocks	**0.914**	0.911	0.913	0.912	0.911	0.901
Phishing_dataset	0.678	0.566	0.576	0.566	**0.769**	0.620
SMK-CAN-187	0.607	0.560	0.613	0.586	**0.762**	0.602
Sonar	0.582	0.588	0.568	0.609	**0.660**	0.610
Spambase	0.527	0.648	0.603	0.653	0.661	**0.687**
Thyroid	0.589	0.589	**0.628**	0.601	0.601	0.608
Titanic	**0.788**	0.715	0.717	0.715	0.777	0.767
Wine	0.848	0.843	0.820	0.865	**0.871**	0.798
Zoo	0.688	0.848	0.360	**0.850**	0.628	0.626

on all evaluated classifiers, SVM, KNN,NB and MLP. Hence, we can reach the classification performance of supervised feature selection methods without using the class labels. We also examined the difference between the moment-based



(a) V20: skewness= 0.61525, variance= 0.07863, IG: 0.00000

(b) V17: skewness= 0.64913, variance= 0.07428, IG: 0.00000

(c) V21: skewness= -0.08424, variance= 0.07348, IG: 0.04589

FIGURE 3. Histograms of features having the highest variance in the sonar dataset.

TABLE 8. Average MSE of KNNR on 10 datasets using skewness and variance-based feature selection, k = [1,30].

Dataset	Skewness	Variance
Automobile	**0.203**	0.228
Battery RUL	55.728	**42.412**
Cargo	0.915	**0.762**
Co2 emissions Canada	13.456	**11.569**
Communities	**0.203**	0.228
Forest fires	**55.057**	56.318
House Price	**20.036**	31.627
Sgemm product	**7.959**	155.042
Song popularity	**21.212**	22.036
Super Conductivity	17.839	**17.468**

TABLE 9. Average MSE of linear regression on 10 data sets using skewness and variance-based feature selection, k = [1,30].

Dataset	Skewness	Variance
Automobile	**0.211**	0.213
Battery RUL	221.252	**89.350**
Cargo	0.824	**0.772**
Co2 emissions Canada	20.103	**19.846**
Communities	**0.211**	0.213
Forest fires	**42.389**	42.491
House Price	31.889	**16.823**
Sgemm product	**1.693**	145.215
Song popularity	21.744	**21.736**
Super Conductivity	31.411	**27.172**

TABLE 10. Average MSE of RFR on 10 data sets using skewness and variance-based feature selection, k = [1,30].

Dataset	**Skewness**	**Variance**
Automobile	**0.193**	0.228
Battery RUL	50.738	**31.454**
Cargo	0.92	**0.672**
Co2 emissions Canada	12.259	**9.053**
Communities	**0.193**	0.227
Forest fires	59.013	**53.467**
House Price	**16.823**	31.605
Sgemm product	**1.893**	145.867
Song popularity	**18.548**	19.196
Super Conductivity	16.545	**16.381**

metrics and the Laplacian Score, an unsupervised filter-based FS algorithm, using the same test. The results show that the difference between these methods is not statistically significant. It is also noteworthy that for each classifier, there were several datasets, where skewness outperformed variance. The obtained results imply that some unknown feature characteristics may determine the best feature subset for a given dataset. It is still an open problem how to choose the best feature selection method for a given data classification or regression task. For example, as shown in table 4, even a simple feature ranking method using variance obtains better results than all other methods on the Breast-Cancer-Wisconsin dataset. Besides, a paired two-tailed t-test was conducted between the best accuracy obtained by one of the supervised methods and one of the unsupervised methods in each dataset. The experimental results, in conjunction with Theorem 1, show that there is no statistically significant difference between the supervised and the unsupervised approaches, implying again that there is no apparent advantage for using class labels in the feature selection process.

To summarize the experimental results above, we have several statistically significant findings: 1) Skewness can serve as an effective filter-based unsupervised feature selection method 2) Variance and skewness-based unsupervised feature selection metrics can reach the classification performance of filter-based supervised methods such as FCBF, MRMR and MIM. 3) Variance and skewness-based feature selection metrics can also reach the accuracy of the unsupervised Laplacian score with less computational efforts.

D. REGRESSION

Since the variance and skewness-based feature selection methods generated comparable results to those of

TABLE 11. Run time in seconds of the evaluated methods for classification datasets.

Dataset	Skewness	Variance-MinMax	Variance-Standard	FCBF	MIM	MRMR	LS
Arrhythmia	**0.001**	0.008	0.009	1.477	88.704	55.295	0.050
Bank	0.012	**0.008**	0.014	5.118	15.744	3.585	10.569
Biodeg	**0.003**	0.007	0.006	0.351	2.277	1.952	0.215
Breast-cancer-wisconsin	**0.002**	0.005	0.018	0.109	0.106	0.107	0.091
Car	**0.002**	0.015	0.010	0.070	0.099	0.098	0.497
Chess	0.081	**0.005**	0.005	0.544	6.084	4.485	1.421
CNAE-9	**0.003**	0.021	0.042	117.331	35.221	35.299	0.459
CTGs	**0.002**	0.004	0.008	0.817	3.547	3.323	0.824
Darwin	**0.007**	0.010	0.010	1.233	113.009	0.626	0.023
Diabetes	**0.001**	0.003	0.003	0.048	0.119	0.117	0.119
Digits	**0.002**	0.004	0.005	1.631	8.773	1.265	25.396
Dis	**0.004**	0.004	0.005	0.714	4.460	0.795	1.363
Dry_Bean	**0.002**	0.005	0.007	2.167	11.242	1.618	34.396
EEG	**0.002**	0.005	0.006	0.995	4.130	4.081	33.570
Glass	**0.003**	0.005	0.009	0.022	0.050	0.050	0.013
Heart stalog	0.004	0.004	**0.003**	0.046	0.095	0.095	0.018
Heart-c	0.005	0.006	**0.003**	0.051	0.105	0.105	0.022
Hepatitis	**0.001**	0.003	0.003	0.030	0.108	0.106	0.010
Ionosphere	**0.003**	0.004	0.003	0.180	0.825	0.807	0.028
Lymphography	**0.012**	0.014	0.015	0.050	0.085	0.084	0.009
Musk	**0.003**	0.019	0.036	6.384	63.682	63.811	9.662
Page-blocks	**0.002**	0.005	0.006	0.558	1.494	0.499	4.982
Phishing	**0.001**	0.005	0.008	3.521	13.208	13.420	18.339
SMK-CAN-187	**0.014**	0.294	0.338	42.325	324.85	318.395	35.245
Sonar	**0.002**	0.007	0.006	0.161	1.091	1.066	0.014
Spambase	**0.001**	0.006	0.007	1.732	14.484	13.912	4.218
Thyroid	**0.004**	0.005	0.007	2.149	18.323	1.596	16.501
Titanic	0.005	0.004	**0.003**	0.084	0.163	0.094	0.124
Wine	**0.002**	0.003	0.003	0.028	0.087	0.086	0.010
Zoo	**0.004**	0.016	0.026	0.029	0.050	0.050	0.006
AVG	**0.006**	**0.017**	**0.021**	**6.332**	**24.407**	**17.561**	**6.606**

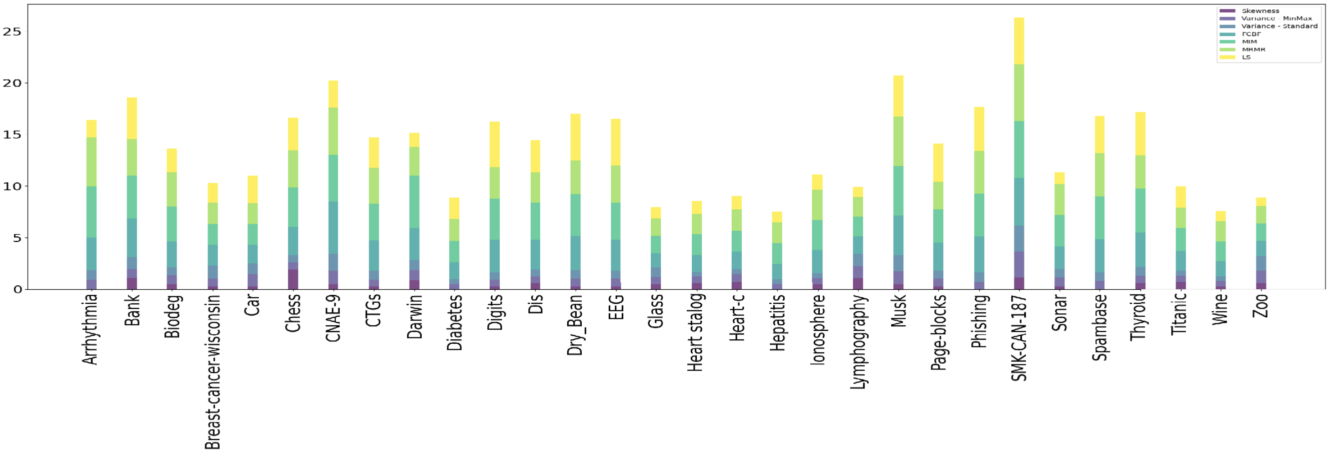


FIGURE 4. Run time (in log of ms) of the feature selection metrics in each classification dataset.

well-established filter-based supervised methods such as FCBF, MRMR, and MIM, it is worthwhile exploring the efficacy of these unsupervised methods for the regression task, where they may also offer a more practical and computationally efficient alternative to commonly utilized supervised methods. This sub-section presents our experimental results for several popular regression models.

E. EXPERIMENTAL SETTINGS FOR REGRESSION TASKS

As skewness and variance-based feature selection methods employ the filter approach, their effectiveness may vary across different regression models. In our evaluation experiments, we measured the Mean squared error (MSE) of three widely utilized regression classifiers, specifically the k-Nearest neighbors Regressor (KNNR), Linear Regression (LR), and Random Forest Regressor (RFR) without incorporating feature selection mechanisms. Our analysis employed 10-fold cross-validation and Min-Max normalization, and we reported the average MSE for each regressor over the range of top-1 to top-30 selected features, with increments of 1. If the number of original features in the dataset was less than 30, the highest value of selected features was limited to the number of original features.

TABLE 12. Run time in seconds of skewness versus variance metrics in each regression dataset.

Dataset	Skewness - No Normalization	Variance-MinMax	Variance-Standard
Automobile	**0.001**	0.003	0.004
Battery_RUL	**0.002**	0.004	0.005
Cargo	**0.005**	0.009	0.011
CO2 Emissions_Canada	**0.002**	0.004	0.005
Communities	**0.004**	0.007	0.010
Forestfires	**0.001**	0.003	0.004
House_price_multifeatures	**0.077**	0.093	0.285
OnlineNewsPopularity	0.032	**0.020**	0.044
sgemm_product	0.063	**0.055**	0.089
Song_data	**0.003**	0.006	0.008
SuperConductivity	**0.022**	0.031	0.037
AVG	**0.019**	**0.021**	**0.046**

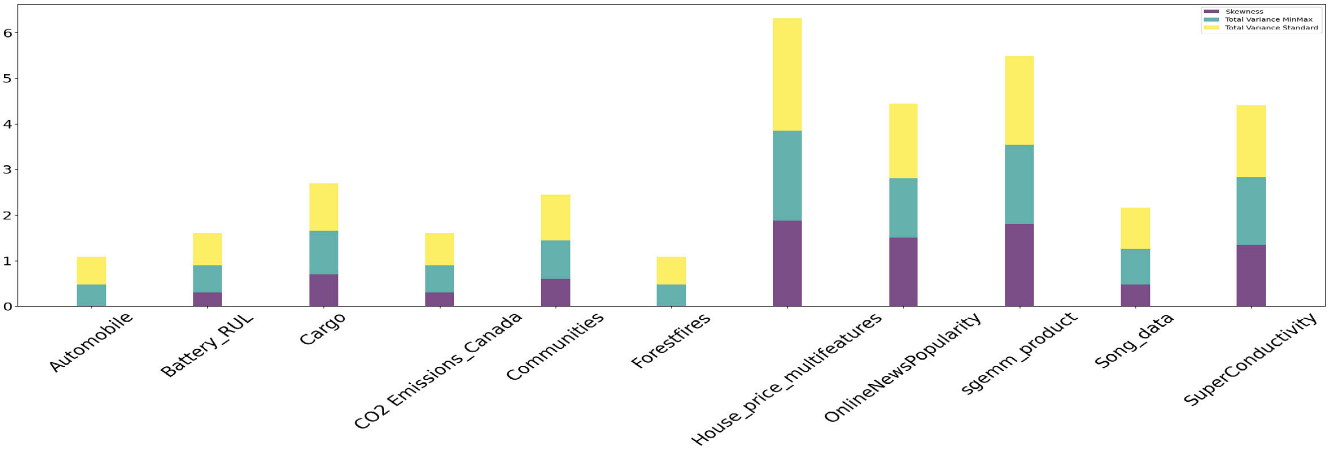


FIGURE 5. Run time (in log of ms) of the feature selection metrics in each regression dataset.

1) MEAN SQUARED ERROR RESULTS

Tables 8-10 show the results of the unsupervised moment-based feature selection methods on 10 benchmark datasets using three different regressors.

We compared the Mean Squared Error (MSE) outcomes of unsupervised feature selection methods based on variance and skewness for regression tasks using the Wilcoxon signed-rank test. Our results demonstrate that the difference between the performance of the evaluated classifiers - KNNR, Linear Regression, and Random Forest Regressor - when using skewness and variance-based feature selection methods is not statistically significant (alpha = 0.05). The main takeaway from this analysis is that in regression tasks, feature ranking using the third statistical moment is as effective as the commonly utilized variance-based method.

F. RUN TIME

In addition to the predictive effectiveness of the second and the third statistical moment-based feature ranking algorithms, we also investigated their runtime performance and compared it with well-established filter-based supervised and unsupervised methods such as FCBF, MRMR, MIM, and Laplacian Score.

Table 11 shows the run time of the evaluated feature ranking methods for classification. Normalization time is included in the feature variance calculation while the skewness calculation does not require normalization. Times are shown in seconds. The code was run on Dell computer with 16 GB RAM and Intel core i7-7600U CPU, 2.80GHz 2.90 GHz.

Table 12 shows relatively short run times of skewness-based FS for unsupervised feature selection in regression-related datasets. Specifically, There are two datasets that show longer skewness calculation time than variance with MinMax normalization. it can be seen that Standard normalization takes the longest time to compute for all datasets. Figure 5 presents the run time charts of the evaluated feature selection metrics in each dataset.

Our analysis reveals that variance and skewness-based feature selection methods are significantly less computationally expensive compared to other supervised and unsupervised filter-based methods, demonstrating the potential of statistical-based techniques to offer practical and computationally efficient solutions to feature selection. Furthermore, skewness is insensitive to data normalization, ensuring consistent feature ranking regardless of the normalization method. In contrast, feature ranking techniques relying on variance may yield different feature rankings depending on normalization methods and require additional run time (by 283% on average using min-max normalization or by 350% using standard scalar normalization).

VIII. CONCLUSION AND FUTURE WORK

Unsupervised feature selection methods have an advantage in various applications due to their ability to select features efficiently from high-dimensional and unlabeled data. In this paper, we have evaluated unsupervised feature ranking and selection metrics based on the second and third statistical moments of each feature (variance and skewness). To the best of our knowledge, this is the first attempt to evaluate skewness as an unsupervised feature scoring metric. The performance of these metrics was examined on 40 benchmark datasets using four different classifiers including linear Support Vector Machine, K-nearest neighbors, Naive Bayes and MLP neural network and three different regression algorithms including k-Nearest neighbors Regressor, Linear Regression, and Random Forest Regressor. Furthermore, the statistical based FS methods were compared to the state-of-the-art unsupervised filter-based feature selection method, Laplacian score (LS), as well as to popular supervised filter-based methods: Fast Correlation Based Filter (FCBF),Minimum Redundancy Maximum Relevance (MRMR) and Mutual Information Maximization (MIM) on classification tasks. The experimental results show that variance and skewness can be used to select a subset of features as effectively as FCBF, MRMR, MIM and LS without the need to use class labels. The results of a theoretical analysis summarized in Theorem 1 provide a statistical explanation of why variance and skewness may be indicative of the feature's actual discriminative power. Variance and skewness-based FS does not require discretization of continuous variables, like entropy-based methods. As opposed to variance-based feature selection methods, skewness-based methods produce stable and consistent rankings, as skewness values remain stable across scaling or normalization of datasets. This implies that utilizing skewness-based feature ranking methods for feature selection offers a reliable and robust approach to generating accurate and stable feature rankings for classification and regression tasks.

Possible directions for future work include extending the experiments using additional learning algorithms and bigger datasets. The focus should be on choosing the most suitable feature scoring method based on the specific characteristics of a given dataset. Another extension would be to use variance, skewness, and their combinations with other feature scoring metrics to choose features in semi-supervised or streaming settings. Also, the effectiveness of positive, negative, and absolute skewness values needs further exploration. Moreover, unsupervised feature metrics may be applied with clustering algorithms, such as k-means.

REFERENCES

[1] M. Belkin, D. Hsu, and J. Xu, "Two models of double descent for weak features," *SIAM J. Math. Data Sci.*, vol. 2, no. 4, pp. 1167–1180, Jan. 2020.

[2] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: A survey of more than two decades of research," *Knowl. Inf. Syst.*, vol. 66, no. 3, pp. 1575–1637, Mar. 2024.

[3] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A survey on semi-supervised feature selection methods," *Pattern Recognit.*, vol. 64, pp. 141–158, Apr. 2017.

[4] S. Alelyani, *On Feature Selection Stability: A Data Perspective*. Princeton, NJ, USA: Citeseer, 2013.

[5] A. U. Haq, D. Zhang, H. Peng, and S. U. Rahman, "Combining multiple feature-ranking techniques and clustering of variables for feature selection," *IEEE Access*, vol. 7, pp. 151482–151492, 2019.

[6] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. Int. Conf. Mach. Learn.*, 2003, pp. 856–863.

[7] D. Bensasson, "Mitochondrial pseudogenes: Evolution's misplaced witnesses," *Trends Ecol. Evol.*, vol. 16, no. 6, pp. 314–321, Jun. 2001.

[8] R. Johnson and T. Zhang, "Semi-supervised convolutional neural networks for text categorization via region embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 919–927.

[9] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, 2014, pp. 584–599.

[10] K. S. Aboody, A. Brown, N. G. Rainov, K. A. Bower, S. Liu, W. Yang, J. E. Small, U. Herrlinger, V. Ourednik, P. M. Black, X. O. Breakefield, and E. Y. Snyder, "Neural stem cells display extensive tropism for pathology in adult brain: Evidence from intracranial gliomas," *Proc. Nat. Acad. Sci. USA*, vol. 97, no. 23, pp. 12846–12851, Nov. 2000.

[11] J. Novakovic, P. Strbac, and D. Bulatovic, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugoslav J. Oper. Res.*, vol. 21, no. 1, pp. 119–135, 2011.

[12] W. Duch, T. Wieczorek, J. Biesiada, and M. Blachnik, "Comparison of feature ranking methods based on information entropy," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Sep. 2004, pp. 1415–1419.

[13] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[14] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 27–66, 2012.

[15] D. D. Lewis, "Feature selection and feature extract ion for text categorization," in *Proc. Speech Natural Lang., Workshop Held Harriman*, New York, NY, USA, Feb. 1992, pp. 1–6.

[16] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[17] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, pp. 1415–1438, Mar. 2003.

[18] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Feature subset selection problem using wrapper approach in supervised learning," *Int. J. Comput. Appl.*, vol. 1, no. 7, pp. 13–17, Feb. 2010.

[19] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," in *Data Classification: Algorithms and Applications*, 2014, p. 37.

[20] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *J. Mach. Learn. Res.*, vol. 5, pp. 845–889, Aug. 2004.

[21] P. Zhu, W. Zhu, Q. Hu, C. Zhang, and W. Zuo, "Subspace clustering guided unsupervised feature selection," *Pattern Recognit.*, vol. 66, pp. 364–374, Jun. 2017.

[22] R. Xu, S. Damelin, B. Nadler, and D. C. Wunsch, "Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps," *Artif. Intell. Med.*, vol. 48, nos. 2–3, pp. 91–98, Feb. 2010.

[23] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[24] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 507–514.

[25] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2010, pp. 333–342.

[26] S. Sharifipour, H. Fayyazi, M. Sabokrou, and E. Adeli, "Unsupervised feature ranking and selection based on autoencoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3172–3176.

[27] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Proc. 7th IEEE Int. Conf. Tools With Artif. Intell.*, Nov. 1995, pp. 388–391.

[28] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowl. Inf. Syst.*, vol. 34, no. 3, pp. 483–519, Mar. 2013.

[29] X.-F. Song, Y. Zhang, D.-W. Gong, and X.-Z. Gao, "A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 9573–9586, Sep. 2022.

[30] M. Petković, S. Džeroski, and D. Kocev, "Feature ranking for semi-supervised learning," *Mach. Learn.*, vol. 112, no. 11, pp. 4379–4408, Nov. 2023.

[31] R. A. Groeneveld and G. Meeden, "Measuring skewness and kurtosis," *J. Roy. Stat. Soc., Ser. D, Statistician*, vol. 33, no. 4, p. 391, Dec. 1984.

[32] A. Rosenfeld, R. Illuz, D. Gottesman, and M. Last, "Using discretization for extending the set of predictive features," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 1, p. 7, Dec. 2018.

[33] D. P. Doane and L. E. Seward, "Measuring skewness: A forgotten statistic?" *J. Statist. Educ.*, vol. 19, no. 2, Jul. 2011.

[34] D. V. Sridhar, E. B. Bartlett, and R. C. Seagrave, "Information theoretic subset selection for neural network models," *Comput. Chem. Eng.*, vol. 22, nos. 4–5, pp. 613–626, Jan. 1998.

[35] T. F. Chan, G. H. Golub, and R. J. LeVeque, "Algorithms for computing the sample variance: Analysis and recommendations," *Amer. Statistician*, vol. 37, no. 3, p. 242, Aug. 1983.

[36] P. P. Pebay, "Formulas for robust, one-pass parallel computation of covariances and arbitrary-order statistical moments," Sandia Nat. Laboratories, Tech. Rep., 2008.

[37] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic, "On mutual information maximization for representation learning," 2019, *arXiv:1907.13625*.

[38] D. Newman, S. Hettich, C. Blake, C. Merz. (1988). *UCI Repository of Machine Learning Databases*. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[39] G. D. Kader and M. Perry, "Variability for categorical variables," *J. Statist. Educ.*, vol. 15, no. 2, Jul. 2007.

[40] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, 2018.

[41] 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," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[42] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in: *Proc. IJCAI*, vol. 93. Princeton, NJ, USA: Citeseer, 1993, pp. 1022–1029.

YAEL HOCHMA received the bachelor's degree in software and information systems engineering and the M.Sc. degree in data science and business intelligence from the Ben-Gurion University of the Negev, Israel, in 2018 and 2019, respectively, where she is currently pursuing the Ph.D. degree. She is also a Data Scientist at Intuit Company, concentrating on NLP, generative AI, and the evaluation process for different machine learning problems. Her research interests include various machine learning topics, such as feature selection and data streams.

YUVAL FELENDLER received the bachelor's degree in software and information systems engineering from the Ben-Gurion University of the Negev, Israel, in 2023, where he is currently pursuing the master's degree in data science and business intelligence. His research interests include feature selection, deep learning, and medical informatics.

MARK LAST received the Ph.D. degree from Tel Aviv University, Israel, in 2000. He is currently a Full Professor with the Department of Software and Information Systems Engineering and the Founding Director of the Data Science Research Center, Ben-Gurion University of the Negev, Israel, and the Head of the Data Engineering Program. Prior to starting his appointment with the Ben-Gurion University of the Negev, in March 2001, he was a Visiting Assistant Professor with the Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA, from 1999 to 2001. He was also the Head of the Software Engineering Program, from 2009 to 2012. He has published about 220 peer-reviewed papers, two monographs, and 11 edited volumes on data mining, text mining, and cyber security. His main research interests include data mining, cross-lingual text mining, soft computing, cyber intelligence, and medical informatics. He is a Senior Member of the IEEE Computer Society and a Professional Member of the Association for Computing Machinery (ACM). He currently serves as an Action Editor for *Data Mining and Knowledge Discovery* and an Editorial Board Member for *Machine Learning* journal and *ACM Transactions on Intelligent Systems and Technology*.