Journal of Universal Computer Science, vol. 22, no. 6 (2016), 760-781
submitted: 30/11/15, accepted: 28/5/16, appeared: 1/6/16 © J.UCS

A Proposal for Recommendation of Feature Selection Algorithm based on Data Set Characteristics

Saptarsi Goswami
(Institute of Engineering and Management
Kolkata, India
Saptarsi007@gmail.com)

Amlan Chakrabarti
(A. K. Choudhury School of Information Technology
Calcutta University, Kolkata, India
acakcs@caluniv.ac.in)

Basabi Chakraborty
(Faculty of Software and Information Science
Iwate Prefectural University, Takizawa, Japan
basabi@iwate-pu.ac.jp)

Abstract: Feature selection is an important prerequisite of any pattern recognition, machine learning or data mining problem. A lot of algorithms for feature subset selection have been developed so far for reduction of dimensionality of the data set in order to achieve high recognition accuracy with low computational cost. However, some methods or algorithms work well for some of the data sets and perform poorly on others. For any particular data set, it is difficult to find out the most suitable algorithm without some random trial and error process. It seems that the characteristics of the data set might have some effect on the algorithm for feature selection. In this work, the data set characteristics is studied for recommendation of appropriate feature selection algorithm to be used for a particular data set. A new proposal in terms of intra attribute relationship and a measure MVS (multivariate score) has been introduced to quantify and group different data sets on the basis of the data set correlation structure into several categories. The measure is used to group 63 publicly available bench mark data set according to their characteristics. The performance of different feature selection algorithms on different groups of data are then studied by simulation experiments to verify the relationship o f data set characteristics and the feature selection algorithm. The effect of some other data set characteristics has also been studied. Finally a framework of recommendation regarding the choice of proper feature selection algorithm has been indicated.

Key Words: Feature selection algorithm, data set characteristics, correlation structure, multivariate score
Category: E.1, H.0, H.4, M.1

1 Introduction

The area of pattern recognition [Duda, 00] deals with classifying known patterns according to a set of labelled training data in a supervised manner while unsupervised classification or clustering [Agarwal, 01] is one of the most widely used technique for exploratory data analysis in which the unlabelled data set is partitioned into a number of meaningful categories or classes known as clusters. Data mining is a related area in which patern classification techniques are used for knowledge extraction from data generated in various applications like medical and health care, finance and business, market behaviour or social data etc. Pattern recognition comprises of two steps : feature selection and classification. The objective of feature selection is to reduce the dimensionality of the data set by removing irrelevant information in order to achieve well defined classes or clusters by efficient classification algorithms with less memory and low computation time. The success of any classification process depends heavily on the proper choice of feature subset selection algorithm in the initial step.

The area of pattern classification or cluster analysis has a long history of research and a lot of algorithms [Theodoridis, 08], [Everitt, 01] have been developed so far. However, it seems that some algorithms work well for some of the data sets but perform poorly for some others [Michie, 94], no single algorithm performs best for all problems as supported also by *no free lunch* theorem [Wolpert, 97]. Every classification model has some underlying assumptions and if it captures the characteristics of the data to be analyzed, the performance of the model seems to be better. Feature subset selection is the most important step before classification and a lot of algorithms for feature subset selection are also available. But there is no single algorithm which performs uniformly well for all data sets. The selection of the most suitable feature subset selection algorithm for a particular data set depends on several considerations like computational cost, usage of memory, accuracy of the output etc. which mainly are the characteristics of the algorithm. Though it is very important to match the algorithm to the characteristics of the data set to be a candidate for its selection for a particular application, no satisfactory unique criterion for selection of suitable algorithm for feature selection process of a particular data set is available till now.

In this paper we focus on analyzing the characteristics of the available data sets and group the data set according to the similar characteristics. The characteristics of the data set can be expressed by different parameters using general properties like number of samples, number of features, number of classes etc., statistical properties like correlation, skewness or kurtosis or information theory based properties like entropy or mutual information. Here we have studied the correlation structure of the data set, that means interaction of the features or intra feature correlation to characterize the data set. A metric MVS (multivariate score) to quantify the strength of correlation values between features has been proposed and has been used to charaterize a data set. The proposed metric is then used to compute intra feature correlation strength for 63 publicly available bench mark data sets. The data sets are then clustered based on MVS values and four natural clusters or groups are found which are denoted as strong independent, weak independent, weak correlated and strong correlated groups. The presently available feature selection algorithms for supervised and unsupervised classification problems are then used to the publicly available bench mark data sets and the performance of univariate and multivariate algorithm have been studied to establish the relationship between the characteristic of the data set with the type of the algorithm.

We also examined data set characteristics in terms of the other statistical properties like skewness and kurtosis and information theoretic property like entropy and how they can be used for characterization of the feature selection algorithm. The next section represents some related works. The following section contains our proposed study of data set characteristics. Section 4 demonstrates simulation experiments and results for studying the effectiveness of our study while the final section contain the summarization of the work and conclusion.

2 Related Works

The effect of data set characteristics in selection of learning algorithms have been studied in a few research works starting from StatLog [Michie, 94] project. In [Smith, 02], the complexity and the characteristics of the data by considering simple, statistical and information theoretic measures has been assessed and several data mining algorithms are examined according to the data set characteristics. In [Ali, 06], an extensive evaluation of learning algorithms for classification problem has been presented with a description of which algorithm are suited to solve which types of classification problem. In [Lee, 13], a technique for selection of learning algorithm for data mining by clustering with behaviourally similar algorithms has been introduced.

In [Smith-Miles, 08], meta learning i.e., learning about learning algorithm performance first used in this context in [Aha, 92], is used for the selection of classification algorithm. In [Song, 12], an algorithm has been developed for feature extraction of a data set, called meta features, using structural and statistical information of the data set to characterize the data sets and a method of recommendation of the classification algorithms have been developed based on the data set characteristics. In [Wang, 13], a technique for recommending an appropriate feature selection algorithm based on meta features (features that characterizes a data set) of the data set has been reported. In both the works, [Song, 12] and [Wang, 13], the recommendation of the classification or feature selection algorithm for a particular data set is done by identifying similar data sets and ranking the candidate target (classification or feature selection) algorithms according to their performance on the similar algorithms. In another recent work, [Wang, 14] represented the algorithm recommendation method from two aspects: meta features to characterize the learning problem and meta target to characterize the relative performance of the classification algorithm on learning problem. They proposed a generic multitarget algorithm recommendation technique as there might be several algorithms suitable for a particular data set.

The most of the algorithms developed so far for recommendation of algorithm selection for feature selection or classification in pattern recognition or data mining problems can be viewed as a meta learning problem of meta features for meta targets. But for successful meta learning, the ratio of examples of two classes is small at the metalevel for any reasonable number of algorithms to choose from and there are risk of overfitting as the algorithms are similar. For fair comparison and selection of algorithms the choice of meta feature should be uniform for different problems, easy to calculate and has good relevance to the performance of the algorithms. To by pass these problems, in this work characterization of data set based on an integrated single measure has been investigated.

2.1 Multivariate Data Sets and Pattern Recognition Algorithms

Generally a data set which contains instances/samples of a single variable is called a univariate data set while the data sets containing instances of multiple variables are called multivariate data set. Theoretically in multivariate data sets, the variables should be independent of each other. In real world, for multi variable data sets, the variables are not always independent, the variables have mostly linear correlation to varying degree. Thus in practical multivariate data sets, the individual features can be strongly independent, strongly correlated or in between.

The classification techniques can also be classified as univariate techniques or multivariate techniques. As an example, for a Naive Bayes' Classifier [Rish, 01], the underlying assumption is that all attributes are independent, so it can be considered as an univariate technique though it can be applied to classify multidimensional data considering the individual features are independent of each other.

As another example, when K-Means clustering or any distance based clustering method is used, if a typical l2-norm or Euclidean norm is used as distance metric , then it is univariate in nature, but if Mahalanobis distance is used, then the same technique is multivariate. CFS, a correlation based feature selection algorithm considers the interaction among variables. The high correlation of the feature with the class is considered to be the quality of the feature to be included in final feature subset while high feature to feature correlation discourages the feature inclusion in the final subset. The technique is inherently considered to be multivariate though it can be used for feature selection to a strongly independent multivariate data set.

In summary, for both the strongly independent data sets and the strongly correlated data sets the feature selection/classification algorithms of following two types can be applied:

1) Univariate techniques that consider the independence of the variables.

2) Multivariate techniques that takes into account the feature correlation.

In this paper the effect of univariate or multivariate techniques to data set of different categories are studied and the results are summarized.

2.2 Association Between Variables

A univariate distribution is described by its mean and variance, whereas for a multivariate distribution, one of the most important constituent is the correlation or covariance matrix. If X is a multivariate normal distribution with k variables, then the probability density function is determined as $X \sim N(\mu, \Sigma)$ where μ is a k dimensional mean vector comprising of the mean of all the k variables and Σ denotes the corresponding covariance matrix. A multivariate distribution can be approximated in terms of the pairwise correlations, between the variables. We discuss a few important, association techniques in this section. We have focused on numerical variables, both continuous and discrete. Nominal, ordinal variables can be coded based on various schemes as available.

2.2.1 Pearson's Product Moment Correlation Coefficient

Pearson's product moment correlation coefficient between two variables x and y, is given by the following equation:

$$\rho(x, y) = \frac{(cov(x, y))}{\sqrt{(var(x) \times var(y))}} \tag{1}$$

Few underlying assumptions are:

The relationship between x and y is linear.

x and y are normally distributed.

The residuals in the scatter plot are homoscedastic i.e. they are random.

$\rho(x, y)$ has a value between 1 and + 1, the higher the absolute value, the higher will be the strength of the relationship. It is also symmetric, i.e., the correlation coefficient between x and y and correlation coefficient between y and x is same.

2.2.2 Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient is defined Pearson's correlation coefficient between the ranked variables as :

$$\rho = 1 - \frac{6\Sigma_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{2}$$

where d_i is the difference between ranks. Pearson's correlation gives meaningful value, when the variables are related by a linear function. However Spearman's correlation coefficient, works for any monotonic function between the two variables. So it can accommodate nonlinear relationship as well.

Interaction of variables can also be expressed by several other information theory based measures like maximal information coefficient (MIC), maximal information compression index [Johnson, 01], information gain, mutual information, symmetric uncertainty etc. In our study, we have used Pearson's correlation coefficient as a measure of intra feature correlation as it is simple to understand, easy to implement and takes less computational time.

2.3 K Correlation Index

K-correlation index is a measure introduced in [Todeschini, 97], for measuring correlation content of multivariate data. In [Todeschini, 99] this index has been further examined and applied to different chemometric field. It is also used for best subset selection of variables in regression problem. The total quantity of correlation contained in a data set is estimated from eigenvalue distribution obtained from the eigenvalue decomposition of the corresponding correlation matrix.

Let $\lambda_1, \lambda_2 \cdots \lambda_p$ be the set of p eigenvalues obtained by PCA (Principal Component Analysis) of correlation matrix. The K-correlation index is defined as follows:

$$K = \frac{\sum_{m=1}^{p} |EV_m - \frac{1}{p}|}{\frac{2(p-1)}{p}} \quad 0 \le K \le 1 \tag{3}$$

where $EV_m = \frac{\lambda_m}{\sum_{m=1}^{p} \lambda_m}$ is the explained variance from the mth principal component. K-correlation index is 1 when all the variables are correlated and 0 when they are uncorrelated. K-correlation index is a measure of global redundancy of the data set and can be used in variable selection by reducing the number of variables while preserving the global correlation structure of the data. This measure is also used for comparison and evaluation of our proposed technique.

3 Proposed Measure for Analysis of Data Structure

The characteristics of the data set is studied here in terms of intra feature correlation structure of the data set. A measure called Multi Variate Score (MVS) has been proposed to express the total strength of intra feature correlation in a data set to quantify the characteristic of the data set. The measure is based on the total of all possible pairwise feature -feature correlation calculated by Pearson's correlation co-efficient. The procedure for obtaining (MVS) of a data set is explained below:

For n dimensional data set, i,e. for n features, there are $n(n-1)/2$ pairs of features. The correlation coefficient of all pairs of features are to be calculated. The absolute values of the correlation coefficients $C_i, i = 1, \cdots, n(n-1)/2$ are grouped into 10 intervals in the range $(0.0 \leq C_i < 0.1, 0.1 \leq C_i < 0.2, 0.2 \leq C_i < 0.3, \cdots, 0.9 \leq C_i \leq 1.0)$. The relative frequency, the ratio of the number of values to total number of values in each interval is calculated. In other words, histograms are drawn for visualization of the correlation distribution of the pairs of features where absolute correlation values (0 to 1) are divided into 10 intervals. Now the distribution of the correlation values (relative frequencies) of a data set is represented by a 10 dimensional vector.

The representation scheme is explained below with Iris [Fisher, 36] data set. Iris data set has 3 classes (3 varieties of Iris plants) of 50 instances each characterized by 4 features sepal length (SL), sepal width (SW), petal length (PL) and petal width (PW). The feature - feature correlation values calculated by Pearson's correlation coefficient are shown below:

SL-SP = -0.1175698 , SL-PL = 0.8717538, SL-PW = 0.81794111, SW-PL = -0.4284401, SW-PW = -0.3661259, PL-PW = 0.9628654 . The histogram representation of the correlation structure of iris data set is shown in Fig 1. The vector representation of the data set, in terms of the relative frequencies of absolute values of pairwise correlation coefficients in ten intervals, become

$D_{iris} = [0.0, 0.17, 0.0, 0.17, 0.17, 0.00, 0.00, 0.00, 0.33, 0.17]$

3.1 Multi Variate Score

The multi variate score (MVS) of a data set (D) , $(MVS)_D$, is represented by the following:

$$(MVS)_D = \Sigma_{i=1}^{10} W_{1i} \times W_{2i} \times D_i \tag{4}$$

where D_i represents the ith component of the vector representation of the pair wise correlation coefficients, W_{1i} and W_{2i} are the ith components of the weight vector W_1 and W_2 respectively. The weight vector W_1 is introduced to consider the effect of overall distribution of correlation coefficient on a large number of data sets and is to be calculated based on the histogram representation of correlation coefficients of all the data sets. The other weight vector W_2 is used to exponentially increase the weight of each interval of the histogram representation. In our study W_2 is set as $[1, 2, 4, 8, 16, 32, 64, 128, 256, 512]$. The

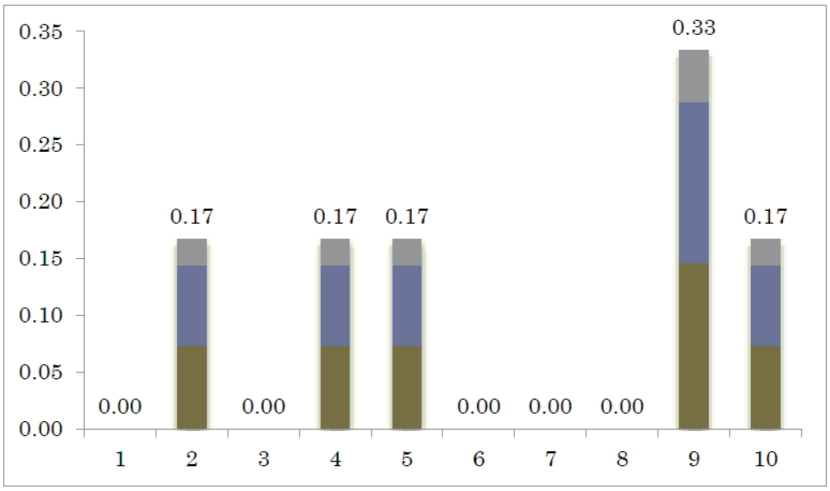

Figure 1: Histogram representation of correlation structure of iris data

Table 1: Calculation for MVS of Iris Data

D_i	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
D_{iris}	0.0	0.17	0.0	0.17	0.17	0.0	0.0	0.0	0.33	0.17
W_1	0.318	1.967	2.814	3.495	4.128	4.711	5.148	5.324	5.684	5.791
W_2	1	2	4	8	16	32	64	128	256	512
$w_{1i} \times w_{2i} \times D_i$	0	0.669	0	4.753	11.23	0	0	0	480.18	504.05

weight W_{1i} is defined as $W_{1i} = log\frac{N}{n_i}$ where N is the sum total number of feature -feature pairs of all the data sets considered, and n_i represents the number of the correlation values that falls on ith interval (total number of interval is 10). Here W_1 is calculated from all the data sets as

$W_1 = [0.318, 1.917, 2.814, 3.495, 4.128, 4.711, 5.148, 5.324, 5.684, 5.791]$

As an example MVS of Iris data set , $(MVS)_{iris}$ came out to be 1000.88 according to Equation 4. This is explained step by step in Table 1.

$(MVS)_D$ is a positive score and it increases with the increase in intra attribute association in terms of correlation coefficient. The range of $(MVS)_D$ can be estimated as follows: Minimum value of MVS, $(MVS)_{min}$ will indicate that absolute value of all correlation coefficients between features of the data set lie between 0 and 0.1 and the maximum value of MVS, $(MVS)_{max}$ will indicate that absolute value of all correlation coefficient lie between 0.9 to 1.0.

Table 2: Data set used

Serial no.	Name of Data set	No. of features.	Sl. no.	Name of Data set	No. of features
1	Appendicitis	7	33	Medlon	500
2	Bands	19	34	Mfeat	649
3	Banknotes	4	35	Magic	10
4	Biodegradation	41	36	Nth	5
5	Bloodtransfusion	4	37	Optdgt	62
6	Braz	857	38	Pageblocks	10
7	Breast Tissue	9	39	Pen	20
8	Bupa	6	40	Phy	78
9	Cleveland	13	41	Pima	8
10	Coli	85	42	Plantleaves	64
11	Contraceptive	9	43	Protein	74
12	Corel	12	44	Saehart	9
13	CTG	32	45	Satellite	36
14	Dermatology	35	46	Satlog	18
15	Digit	255	47	Scene	12
16	Dow	12	48	Seeds	7
17	Ecoli	7	49	Segment	17
18	Fertility	9	50	Shuttle	9
19	Forestcover	53	51	Sonar	60
20	Gender	100	52	Spambase	57
21	Glass	9	53	Specftf	44
22	Heart	13	54	Texture	40
23	Hepatitis	13	55	Twonorm	20
24	ILPD	10	56	Vehicle	18
25	Ionosphere	33	57	Waveform	22
26	Iris	4	58	Wbdc	30
27	Kdigit	783	59	Wine	13
28	Leaves	14	60	Wine Red	11
29	Led	7	61	WineWhite	11
30	LSVT	310	62	Wscon	9
31	Mammogram	5	63	Yeast	8
32	Marketing	13			

Table 3: Categories of Data sets according to MVS

Category of data set	MVS range	Number of data sets	Lowest MVS	Highest MVS
Strong Independent (SI)	≤ 20	19	0.3	15.82
Weak Independent (WI)	$\leq 20 \leq 72.5$	17	23.43	72.5
Weak Correlated (WC)	$\leq 72.5 \leq 150$	9	72.61	130.26
Strong Correlated (SC)	≥ 150	18	188.30	1153.1

4 Simulation Experiments and Results

The proposed method of analysis of data sets has been studied with simulation experiments using 63 publicly available bench mark data sets from [Bache, 13] and [Alcal-Fdez, 11]. The data sets are collected from variety of domains like biology, image processing, medical and text processing. The number of attributes in the data set are also varied from less than 10 to more than 500. Table 2 represents the data sets used.

4.1 Categorization of Data Set According to MVS

The pair wise correlation values of each data set is calculated . The values are used to draw histograms and finally each data set is represented by a 10 dimensional vector. (MVS) for each data set is calculated according to Equation 4. described in the previous section. The value of W_1 is calculated based on all correlation coefficients of the 63 data sets as:

$W_1 = [0.32, 1.97, 2.81, 3.49, 4.13, 4.71, 5.15, 5.32, 5.68, 5.79]$

(MVS) values are partitioned according to PAM (Partitioning around Mediod), implemented as in [Maechler, 12] and cluster validity is checked by Silhouette width [Rousseeuw, 87]. According to Silhouette width, the appropriate number of clusters is around 3 to 5. Considering Silhouette width and visualization of the clusters we decided 4 as the number of appropriate clusters.

The clustering result according to multivariate score is represented in Table 3. The four clusters of data sets are named as strong independent, weak independent, weak correlated and strong correlated according to their (MVS) values. The lowest and highest values of (MVS) in different categories are presented in the table.

The characteristics of the four groups are as follows:

1. Group 1, it is found that the data sets with the minimum and maximum (MVS) in this group do not have any pairs in the top 3 deciles i,e high correlation ranges, so it conforms to the idea of strong independent(SI)
2. Group 2, this group has majority of the values in lower 3 correlation ranges, with some values in 8th and 9th intervals. So it can be considered as 'weak independent '(WI)
3. Group 3, this group is very similar to Group 2. But it is observed that the histogram structure has shifted more towards right compared to Group 2. There are more values in high correlation intervals with some values in the low ranges. So it is called 'weak correlated' (WC).
4. Group 4, it has significant number of correlation coefficients in top three deciles. This is in conformance with the notion of 'strong correlated' (SC).

Figure 2 represents the histogram of the data sets having the lowest and the highest MVS values. The first, second, third and fourth row of the figure represent the histogram of the data sets having lowest and highest MVS value in strong independent, weak independent, weak correlated and strong correlated categories respectively. By examining MVS values (proposed to express total intra feature correlation of the data set) and the feature-feature correlation values (calculated by Pearson's correlation coefficient) , following are observed:

- Strong Independent Group, which has no or minimal Correlation coefficients ≤ 0.6
- Weak Independent Group, which has some correlation coefficients between 0.6 and 0.9 compared to Strong Independent group
- Weak Correlated Group, which has more correlation between 0.6 and 0.9 compared to Weak Independent group
- Strong Correlated Group, which has close to correlation coefficients greater than 0.9.

4.2 Effect of Data Set Category on Type of Algorithm

We have done following simulation experiments to study the effect of the various groups of the data set according to (MVS) values on feature selection algorithms. In our experiment we used strong independent data set and strong correlated data set with both univariate and multivariate feature selection algorithms to select optimal subset of features. As univariate measure (UM), a filter algorithm with information gain theoretic evaluation function has been used. The features

Table 4: Classification accuracies for strong independent data sets

Data set	No. of feat.	Full	CFS	UM1	UM2	UM3
Medlon	500	0.58	0.6	0.6	0.59	0.59
Gender	100	0.63	0.63	0.64	0.65	0.66
Twonorm	20	0.98	0.98	0.85	0.93	0.96
Yeast	8	0.31	0.53	0.53	0.53	0.53
Kdigit	783	0.51	0.78	0.76	0.66	0.57
Digit	255	0.81	0.81	0.66	0.76	0.8
Heart	13	0.83	0.83	0.8	0.83	0.83
Braz	857	0.19	0.65	0.68	0.23	0.21
Spambase	57	0.71	0.86	0.87	0.86	0.7
Bands	19	0.42	0.63	0.65	0.65	0.41
Pima	8	0.75	0.76	0.76	0.75	0.75
Hepatitis	19	0.53	0.82	0.83	0.43	0.46
Plantleaves	64	0.39	0.1	0.18	0.3	0.35
Optdgt	62	0.83	0.91	0.85	0.9	0.91

Table 5: Comparison of execution time for strong independent data set

Data set	MVS value	Time for UM	Time for CFS
Medlon	0.317663	22.5	250.94
Gender	2.056746	1.96	18.29
Twonorm	3.953495	1.74	2.1
Yeast	5.358459	0.13	0.27
Kdigit	6.947401	6.18	136696
Digit	8.193302	5.01	4719.53
Heart	8.398692	0.12	0.25
Braz	9.283189	1.52	5472.09
Spambase	11.11827	2.03	3.24
Bands	12.46301	0.3	0.2
Pima	12.57666	0.08	0.09
Hepatitis	13.7162	0.2	0.17
Plantleaves	13.81892	0.58	1.35
Optdgt	15.82113	7.85	14.52

Table 6: Classification accuracies for strong correlated data sets

Data set	No. of feat.	Full	CFS	UM1	UM2	UM3
Shuttle	9	0.75	0.78	0.77	0.8	0.9
LSVT	310	0.53	0.54	0.53	0.56	0.55
WBDC	30	0.93	0.95	0.94	0.93	0.94
Satlog	18	0.58	0.85	0.81	0.72	0.81
Segment	19	0.79	0.83	0.67	0.79	0.79
Wscon	9	0.96	0.96	0.95	0.96	0.96
Leaves	9	0.64	0.66	0.59	0.62	0.64
Vehicle	18	0.46	0.49	0.41	0.46	0.44
Dow	12	0.7	0.68	0.46	0.56	0.67
Satelite	35	0.1	0.19	0.14	0.12	0.11
Iris	4	0.96	0.96	0.95	0.96	0.96
Seeds	7	0.91	0.91	0.86	0.86	0.88
Texture	40	0.78	0.83	0.8	0.8	0.79

Table 7: Comparison of execution time for strong correlated data set

Data set	MVS value	Time for UM	Time for CFS
Shuttle	188.3026	6.69	3.5
LSVT	278.6828	3.41	404.7
WBDC	312.6853	0.12	0.12
Satlog	324.6994	0.4	0.35
Segment	327.9067	2.51	1.4
Wscon	366.5781	0.13	0.14
Leaves	408.3547	1.68	1.67
Breast Tissue	522.1267	0.08	0.12
Vehicle	538.1621	0.1	0.1
Dow	673.7429	0.19	0.28
Satelite	680.4015	1.17	1.64
iris	854.2605	0.05	0.04
Seeds	1095.818	0.09	0.08
Texture	1153.074	2.12	7.49

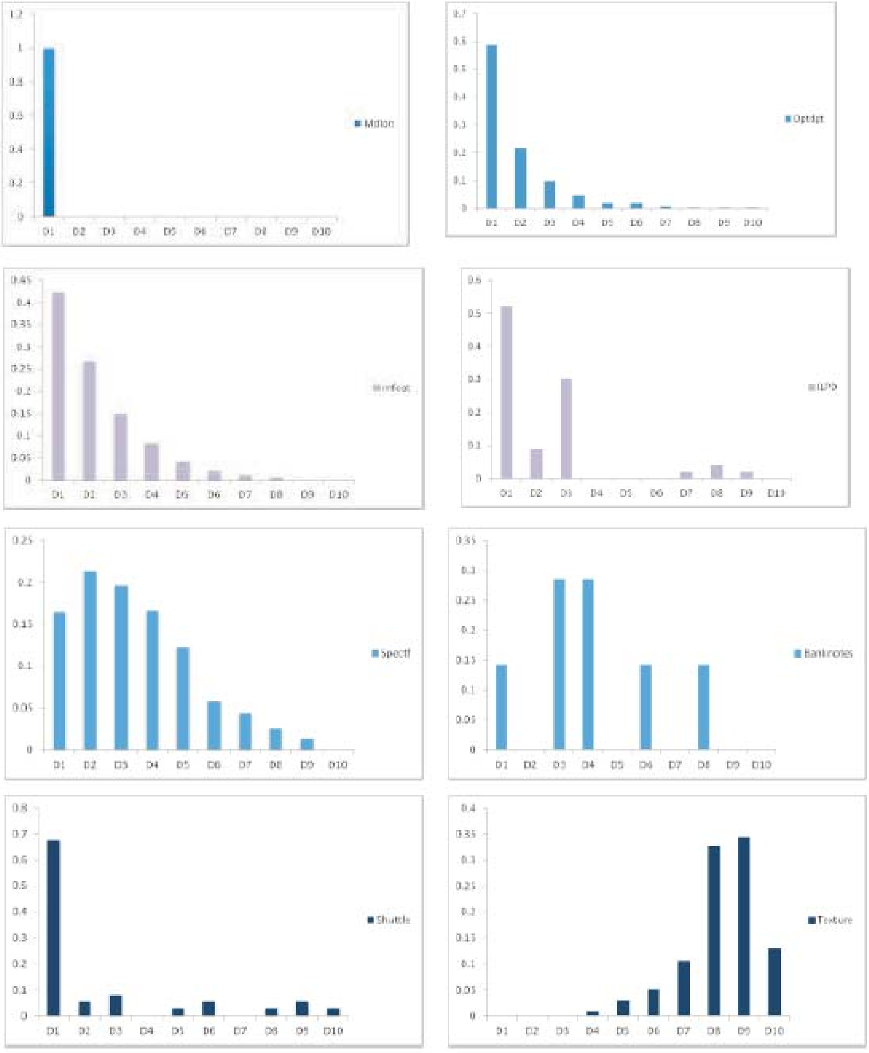

Figure 2: Histogram of lowest and highest (MVS) data set of different categories

are ranked and top 25% or less, 50% and 75% features are retained for evaluation. For multivariate algorithm, CFS (correlation based feature selection) method [Hall, 99] has been used. Naive Bayes' classifier is used for classification. Classification accuracy (average of 25 runs) is used for the evaluation of the performance of the feature selection algorithm.

Table 4 represents the classification accuracies for strong independent data

Table 8: Cardinality of final feature subset

Data set	Original no. of features	No. of features in final subset (UM)	No. of features in final subset (CFS)
Shuttle	9	7	3
LSVT	310	155	24
WBDC	30	23	8
Satlog	18	4	5
Segment	17	5	4
Wscon	9	7	7
Leaves	14	4	8
Breast Tissue	9	7	5
Vehicle	18	12	9
Dow	35	9	10
Satelite	35	3	3
Iris	4	2	2
Seeds	7	2	2
Texture	40	10	10

set (14 data sets out of 18 has been used) classified by our (MVS) score with CFS method and the univariate method. The first column represents the data set, 2nd column represents the number of features in the data set, 3rd, 4th, 5th, 6th and 7th columns represent classification accuracy with full feature set, feature set selected by CFS, feature set ranked by univariate method with top 25% or less (UM1), with top 50% (UM2) , with top 75% (UM3) respectively.

Table 5 represents the execution time in sec. of the two algorithms (univariate (UM) and multivariate CFS) on different data sets. Here 2nd column represents the (MVS) values of the data set.

It has been observed from the results that the univariate method has an average performance gain of 2.5%. The computational time of univariate method is much less compared to multivariate CFS method, specially for high dimensional feature sets such as Kdigit, Braz or Digit.

Table 6 represents the experimental results of classification accuracies with features selected by different algorithms for strong correlated data sets. Table 7 represents comparison of computational time for univariate and CFS method applied to strong correlated data sets.

It has been found that on average, multivariate method produced a performance gain of 1%. However, as shown in Table 8, use of multivariate method for strong correlated data set produces lesser number of features in the final subset, reduction of feature set cardinality is about 10% on average compared to univariate methods. It is also found that the reduction is higher for originally

Table 9: Feature selection for clustering for independent datasets

Category	Dataset Name	univariate	PFA	EVA
SI	Cleveland	0.59	0.59	0.59
SI	Digits	0.58	0.62	0.62
SI	Gender	0.52	0.52	0.52
SI	Mdlon	0.58	0.59	0.58
SI	Spambase	0.76	0.73	0.74
SI	Twonorm	0.97	0.97	0.97
SI	Heart	0.79	0.76	0.79
SI	Pima	0.67	0.67	0.66
SI	Optdgt	0.76	0.52	0.71
SI	Contra	0.43	0.43	0.43
WI	CTG	0.4	0.77	0.45
WI	wqwhite	0.47	0.48	0.46
WI	saehart	0.65	0.68	0.65
WI	ILPD	0.72	0.72	0.72
WI	Sonar	0.56	0.53	0.55
WI	mfeat	0.71	0.84	0.69
WI	biodeg	0.7	0.66	0.66
WI	Glass	0.57	0.53	0.56
WI	Pageblocks	0.91	0.91	0.9
WI	Fertility	0.88	0.88	0.88

higher dimensional data sets. In this case feature set cardinality by using univariate method is set at the number of features for which performance of the classifier (classifier accuracy) is the highest.

We have also used univariate and multivariate algorithms for feature selection in case of unsupervised pattern recognition problems. In Table 9 the classification accuracies of SI and WI i,e independent data sets are presented where feature selection has been done by an entropy based univariate method and two multivaraiate methods marked by PFA [Lu, 07] and EVA [Mitra, 02] in the table. Similarly in Table 10 the classification accuracies of SC and WC i,e. correlated data sets are presented where feature selection has been done with the same algorithms. It can be verified from the results that on average the univariate method works better for independent data sets while multivariate methods are better for correlated data sets.

Table 10: Feature selection for clustering for correlated datasets

Category	Dataset Name	Entropy	PFA	EVA
SC	Appendicities	0.8	0.84	0.8
WC	Darma	0.84	0.76	0.73
SC	Dow	0.56	0.55	0.48
SC	Iris	0.7	0.8	0.96
SC	Magic	0.65	0.66	0.69
WC	Pen	0.57	0.61	0.72
SC	Satimg	0.67	0.59	0.74
SC	Segment	0.66	0.56	0.66
WC	Specfheart	0.79	0.79	0.79
SC	Texture	0.52	0.55	0.52
SC	Vehicle	0.38	0.37	0.39

Table 11: Comparison of MVS with K-index

Dataset	K-index	MVS	Category (MVS)
Appendicites	0.64	614.27	SC
Bands	0.23	12.46	WC
Banknotes	0.49	130.26	WC
Biodegradation	0.52	51.41	WI
Braz	0.52	9.28	SI
Coli	0.45	24.1	WI
Corel	0.661	403.31	SC
Digits	0.57	8.19	SI
Forest	0.16	2.51	SI
Iris	0.64	854.26	SC
Led	0.218	7.56	SI
Marketing	0.31	35.77	WI
Mfeat	0.76	23.42	WI
Magic	0.42	200.363	SC
Pen	0.526	74.84	WC
Plantleaves	0.5	13.81	SI
Shuttle	0.39	188.30	SC
Sonar	0.59	57.47	WI
Waveform	0.48	81.30	WC
Yeast	0.16	5,35	SI

4.3 Comparison with K-correlation Index

We have also grouped the data set characteristics with K correlation Index measure and compared the results with our MVS measure. K-correlation index is focussed on finding overall redundancy where the objective of MVS looks into pair wise correlation in finding redundancy. Table 11 represents the comparison results of some data sets with their MVS values, k-index values and categories according to MVS. Though MVS has a high correlation with K-index, as it assigns exponential weightage on the bins, the demarcations between groups are much more sharper, so easier to identify the characteristics of the data set. Some of the examples are :

- For "shuttle" data set, K-index is 0.39, which shows a moderate correlation while MVS puts this data set in SC (Strong Correlation) group which can be verified by our crosschecking experiments with different feature selection algorithms (produces better results for multivariate algorithm). Similar is the case for "magic" data set.
- Some other data sets like "mfeat", "plantleaves", "Braz", "sonar", "digits" etc.. MVS score shows a lower correlation and classifies these data sets as strong or weak independent while K-index classifies its as highly correlated which contradicts our experiments with algorithms.

Thus it seems that MVS is a better measure in characterizing a data set in terms of correlation structure.

4.4 Effect of Skewness of the Data set

Skewness is a statistical measure of a data set indicating the amount of assymetry in the data distribution. It is actually a measure of the shape of the distribution. It is anticipated that a feature which exhibits very high skewness in its distribution is not suitable for any classification task, supervised and unsupervised. Simulation experiments are conducted on 10 publicly available datasets from UCI, which contained features with high skewness factor. The corresponding symmetrical uncertainty and entropy of the features are also measured. Entropy has been used to measure the information richness of individual features. This can be typically be useful for unsupervised problems as well. As shown in Table 12 , out of the 36 attributes , 30 attributes produce very low symmetrical uncertainty ($\leq$ 0.1) and entropy. Though the experiment was conducted on datasets with target information available so that the symmetrical uncertainty can be measured , the underlying notion is generic enough to be extended to unsupervised problems. This is evident from the low entropy values of these features having high skewness. Hence a further detailed study regarding the shape related properties for features for selection of feature selection algorithms can be useful.

5 Conclusion

In the area of pattern recognition and data mining, lot of algorithms has been developed to partition the data set into known classes or meaningful clusters with an objective to achieve high classification accuracy (clusters with definite boundary in case of unsupervised classification) as well as low computational cost. It has been found that the performance of any algorithm varies over the data sets. Some algorithms perform well for some data sets but cannot give good result for other data sets. It seems that the characteristics of the data set influences the behaviour of the pattern classification algorithm. So, to achieve reasonable performance for a new data set, it can be assumed that the selection of algorithm according to the characteristics of the data set can lead to an improvement of performance. Now data set can be characterized by several parameters.

In this work, the correlation structure of the data set has been considered as a parameter to characterize the data set. The correlation structure of a large number of data sets are examined by a proposed measure. The data sets are then grouped into several categories according to the measure. In simple experiments, the performance of different feature selection algorithms for supervised and unsupervised tasks on different categories of the data sets are examined. Though we have used only feature selection algorithms, classification (supervised) and clustering algorithms also can be used in the same manner for further study. It is found that for strong independent data sets, univariate feature selection methods are well suited. Though the multivariate methods also produce good output in some cases, the computational time is much higher compared to univariate methods. On the other hand, for strong correlated data sets, multivariate method (CFS) produce the feature subset containing lesser number of features than univariate method. So it can be recommended that in case of strong correlated data sets with high dimension, multivariate methods can be used to reduce the feature set in the first step and then univariate method can be used for selecting final subset. In this way, the total computational cost for feature selection process for high dimensional data can be reduced.

Now, the data set characteristics cannot be correctly estimated by only the correlation structure of the data set. We are also investigating other stastistical properties of the data sets like skewness and information theoretic property as entropy so that they can be integrated with correlation structure to define a better metric to characterize a data set. Our preliminary studies with skewness of the data sets indicate that it also a good measure for the first step of reduction of features for a high dimensional data set.

Table 12: Features with high skewness

Dataset	Feature Name	Skewness	Entropy	Symetrical Uncertainty
Pageblocks	meantr	67.39	0.06	0.17
Optdgt	Attr.56	52.97	0	0
Coli	AVRAAUT	39.35	0.01	0
Optdgt	Attr.8	38.9	0.01	0
Coli	PZEILPL	38.28	0	0
Optdgt	Attr.32	33.47	0	0
Coli	AZEILPL	32.98	0	0
Optdgt	Attr.16	32.2	0.02	0.04
Shuttle	A4	31.69	0.02	0
Coli	AWERKT	31.34	0.03	0
Optdgt	Attr.24	30.55	0.01	0
Phy	att41	28	0.17	0
Coli	PVRAAUT	27.5	0.01	0
Biodg	a23	26.8	0.21	0.08
Optdgt	Atr.48	26.3	0.05	0.01
Phy	Att42	24.13	0.16	0
Optdgt	Atr.31	23.4	0.02	0
Phy	att42	24.13	0.16	0
Mamo	BI.RADS	23.1	0.16	0
Shuttle	A6	-21.86	0.17	0
Coli	PWERKT	20.79	0.01	0
Pageblocks	Heights	20.36	0.57	0.13
Coli	PPERSONG	19.9	0.04	0
Pageblocks	area	19.51	0.59	0.1
bands	ESAamperage	18.8	0.04	0
Coli	AWAOREG	18.42	0.03	0
Phy	att40	18.3	0.18	0
Ecoli	Chg	18.2	0.02	0
Coli	ABESAUT	17.6	0.06	0
Coli	ABESAUT	17.5	0.04	0
Segments	Hedge.sd	16.9	0.3	0.07
Coli	APLEZIER	17	0.03	0
Coli	PINBOED	16.3	0.06	0
Coli	PWAOREG	16.2	0.03	0
Optdgt	Atr.40	15.8	0.06	0.02
Optdgt	Atr.47	15.8	0.08	0.01

Acknowledgement

The third author Basabi Chakraborty would like to acknowledge that this work was partly supported by JSPS KAKENHI Grant Number 26280090.

References

[Agarwal, 01] Aggarwal, C.C.: Data Clustering: Algorithms and Applications, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2013.

[Aha, 92] Aha, D.: Generalizing from Case studies* A case study, Proc. of 9th International Conference on Machine Learning", pp. 1-10, 1992.

[Alcal-Fdez, 11] Alcal-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., Garca, S., Snchez, L., Herrera, F.: KEEL Data-Mining Soft-ware Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework, Journal of Multiple-Valued Logic and Soft Computing Vol. 17, No. 2-3, pp. 255-287, 2011.

[Ali, 06] Ali, S., Smith, K.A.: On learning algorithm selection for classification, Applied Soft Computing, Vol. 6, No. 2, pp. 119–138, 2006.

[Bache, 13] Bache, K., Lichman, M.: UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

[Duda, 00] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, Wiley & Sons, Inc., New York, 2nd Edition, 2001.

[Everitt, 01] Everitt, B.S., Landau, S., Leese, M.: Cluster Analysis, Arnold Publishers, London, 4th Edition, 2001.

[Fisher, 36] Fisher, R.A.: The Use of Multiple Measurements in Axonomic Problems, Annals of Eugenics, Vol 7, pp. 179-188, 1936.

[Hall, 99] Hall, A.M.: Correlation-based feature selection for machine learning, Dissertation for Ph. D, The University of Waikato,NewZealand, 1999.

[Johnson, 01] Johnson, R.A., Wichern,D.W.: Applied multivariate statistical analysis, 5th Edition, Englewood Cliffs, NJ: Prentice hall, 2001.

[Lee, 13] Lee, J.W., Giraud-Carrier, C.G.: Automatic selection of classification learning algorithms for data mining practitioners, Intell. Data Anal. Vol. 17 (4), pp. 665-678, 2013.

[Lu, 07] Lu, Y., Cohen, I., Zhao, X.S., Tian, Q.: Feature selection using Principal feature analysis, Proc. of 15th International conference on Multimedia, ACM, pp. 301-304, 2007.

[Maechler, 12] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.: Cluster: Cluster Analysis Basics and Extensions, R package version 1.14.3. 2012.

[Michie, 94] Michie, D., Spiegelhalter, D.J., Taylor, C.C.: Machine Learning, Neural and Statistical Classification, Ellis Horwood, New York, 1994.

[Mitra, 02] Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity, IEEE Trans. on PAMI, Vol. 24, No. 3, pp. 301-312, 2002.

[Rish, 01] Rish, I.: An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence,(available online: PDF (http://www.research.ibm.com/people/r/rish/) papers/RC22230.pdf)

[Rousseeuw, 87] Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, Vol.20, pp. 53-65,1987.

[Smith, 02] Smith, K. A.: Matching Datamining algorithm suitability to data characteristics using a self organizing map, in Hybrid Information Systems, Physica-Verlag, pp. 169-180,2002.

[Smith-Miles, 08] Smith-Miles, K. A.: Cross-disciplinary perspectives on meta-learning for algorithm selection, ACM Computing Surveys, Vol. 41, No. 1, pp. 1-25, 2008.

[Song, 12] Song, Q.B., Wang, G.T., Wang, C.: Automatic recommendation of classification algorithms based on data set characteristics, Pattern Recognition, Vol. 45, No. 7, pp. 2672-2689, 2012.

[Theodoridis, 08] Theodoridis, S., Koutroumbas, K.: Pattern Recognition, Academic Press, 4th Edition, 2008.

[Todeschini, 97] Todeschini, R.: Data correlation, number of significant principal components and shape of molecules. The K correlation index, Analytica Chimca Acta Vol. 348, pp. 419-430, 1997.

[Todeschini, 99] Todeschini, R.: The K correlation index: theory development and its application in chemometrics, Chemometrics and Intelligent Laboratory Systems, Vol 46. Issue 1, pp. 13-29, 1999.

[Wang, 13] Wang, G.T.: A feature subset selection algorithm automatic recommendation method, Journal of Artificial Intelligence Research, Vol. 47, pp. 1-34, 2013.

[Wang, 14] Wang, G.T.: A Generic Multilabel Learning-Based Classification Algorithm Recommendation Method, ACM Tran. on Knowledge DIscovery from Data, Vol. 9, No. 1, Article 7, 2014.

[Wolpert, 97] Wolpert, W., Macready, W.G.: No Free Lunch Theorem for Optimization, IEEE Trans. Evolutionary Computation, Vol.1, No. 1, pp. 67-82, 1997.