A Novel Feature Selection Measure Partnership-Gain

https://doi.org/10.3991/ijoe.v15i04.9831

Mostafa A. Salama (✉)
British University in Egypt (BUE), Cairo, Egypt
mostafa.salama@bue.edu.eg

Ghada Hassan
Ain Shams University, Cairo, Egypt
British University in Egypt (BUE), Cairo, Egypt

Abstract—Multivariate feature selection techniques search for the optimal features subset to reduce the dimensionality and hence the complexity of a classification task. Statistical feature selection techniques measure the mutual correlation between features well as the correlation of each feature to the tar- get feature. However, adding a feature to a feature subset could deteriorate the classification accuracy even though this feature positively correlates to the target class. Although most of existing feature ranking/selection techniques consider the interdependency between features, the nature of interaction be- tween features in relationship to the classification problem is still not well investigated. This study proposes a technique for forward feature selection that calculates the novel measure Partnership-Gain to select a subset of features whose partnership constructively correlates to the target feature classification. Comparative analysis to other well-known techniques shows that the proposed technique has either an enhanced or a comparable classification accuracy on the datasets studied. We present a visualization of the degree and direction of the proposed measure of features' partnerships for a better understanding of the measure's nature.

Keywords—Feature Selection, Interdependency between features, Classification

1 Introduction

The dimensionality reduction of real-life datasets decreases the cost of measuring, storing and processing extra non-useful data features. Feature Selection is a technique of filtering features that are irrelevant or redundant to the target classification problem. Techniques for feature selection are classified into the wrapper and filter techniques. Filter techniques model is independent of the classification algorithm and it evaluates the features based on their statistical relevance to the target feature [4]. Wrapper techniques model utilizes an inductive classifier to evaluate subsets of features based on their discriminative power [6]. To avoid the massive time needed to test all possible combinations of features, wrapper techniques use computationally accepted greedy

search strategies like forward and backward methods in the selection. The optimal sub-set of features is the minimal set of features of high correlation to target class label and low interdependency/redundancy between features.

Feature selection techniques are further classified as either univariate, which are techniques that only consider the relevancy between each feature and the target class, or multivariate, which are techniques that consider, in addition, the inter-correlation between features. Although the correlation between a sub-set of features is high, the combination of these subset in the same data set is not necessarily leads to high classification accuracy percentage. For example, local minima peaks may appear in forward feature selection techniques resulting from adding a feature to another set of features even though this feature, independently, has a high correlation to the target class. Other techniques of multivariate feature selection also do not address this problem. Statistical methods like information gain calculate the correlation between features to remove redundancy. While ensemble bagging, boosting and staking methods like random forest may be prone to over-fitting due to overemphasizing noise in the input dataset [21, 25].

This work proposes a Partnership-Gain measure that selects features to include in the feature subset based on how the features' partnership contributes to the classification task at hand. The proposed measure shows whether the correlation between features has a constructive, destructive or neutral effect on the classification problem. Constructive correlation is indicated when the classification accuracy using a specific feature subset is higher than the accuracy of using any one single feature from the subset. Destructive correlation is indicated when the accuracy of a feature set is lower than the accuracy of using one of the features in the subset. The features determined to have a constructive/destructive effect are said to be in a positive/negative partnership with respect to the particular classification task, and hence the decision to include/not include features in the feature subset. A neutral partnership is an indication of the redundancy of features where the classification accuracy is not affected by having them all or only one of them in the feature subset. The competency of the proposed technique is tested against other feature selection state-of-the-art techniques. The techniques used in the comparison study include: feature selection algorithm which is relief, forward feature selection algorithms based on feature ranking methods which are information gain and Chi-Merge, and wrapper support vector machine classifier that is used as feature evaluation criterion for selecting features [18]. Experiments are conducted to compare the classification accuracy resulting from using the proposed technique against the classification accuracy of other feature selection techniques on seven benchmark datasets using Bayesian belief network as an evaluating classifier. The results show a good performance of the proposed technique relevant to the existing techniques. Results are also helpful to visualize the relation between the selected features and the significance of the feature's partnership with respect to a classification problem.

The rest of this paper is organized as follows: The reviews and critics of related work. A presentation of the stated hypothesis in this work, the proposed Partnership-Gain measure, and the feature subset selection algorithm. The experimentally, the evaluation of the importance of the proposed technique is presented, then the classification results of the technique in comparison to other well-known feature selection techniques

on a number of benchmark classification tasks. Finally, a visualization of the partnership gain measure between selected features is presented.

2 Background

Current selection techniques are classified into univariate methods that detect the patterns existing in the values of a single feature with respect to the target class, and multivariate methods that study the interconnection between features in addition to the correlation of each feature to the target class [24]. Information gain is a univariate technique that measures the mutual information between a single feature and the target class [26]. Mutual information MI (T; F) measures how much the uncertainty (entropy: marginal probability distribution) of a target class T is reduced if a feature F has been observed [25]. Measuring the relevance between features using information gain is biased towards features with higher value ranges [19]. Chi-square and Chi-merge methods measure the dependency of an attribute on the target class labels according to the variance of the values of this attribute [5]. These univariate methods ignore the contribution of the multivariate patterns distributed among features in the classification of instances.

The multivariate techniques consider a group of features rather than a single feature in the search for patterns required in the classification task in order to detect the least redundant and most relevant set of features. The correlation coefficient measure between two feature vectors can be a linear or non-linear measure. Real life problems usually require non-linear measures like information gain. The relevance between two features can be categorized as strong, weak or not relevant. A Strong relevance indicates either the existence of high redundancy between features or their information is complementary. The structure of detecting the interaction between features in a group of well-known multivariate feature selection techniques is discussed here as follows:

- Relief-F is an iterative weighting algorithm [11] that updates the features' weight in each iteration based on randomly selected instances. The algorithm changes the weight of each feature according to its Euclidean distance from the nearest instances existing in different target classes. Accordingly, Relief fails to deal with the outliers instances [16].
- The Joint Mutual Information (JMI) is another example of filter multivariate techniques. Joint probability distribution is written as P(T | A,B,C ,..), which is the probability of the target class T given features A, B, C , ... The sum of the pairwise joint mutual information between features with respect to the target features is calculated. The selected optimal subset of features is the set that maximizes the JMI value. Definition of the JMI means that adding a feature to the pre-calculated features will never decrease the JMI value [17]. Selection of this subset of features requires an exhaustive search for calculating the JMI estimates for all the possible feature subsets. Bayesian network (BN) is a directed acyclic graphical representation of the dependencies among features. The dependency between features X and Y is given by equation (1):

$$MI(X:Y)=\sigma_{(x,y)}\, p(x,y)^{*} [\![\log]\!]_2 \ (p(x,y))/(p(x)p(y)) \quad (1)$$

- Such that x is all possible values of feature X, and so is y, while p(x,y) is the joint probability that the value of X is x and the value of Y is y. The target class is represented by a node in BN, and any arc from this node to other nodes representing features means that target class is dependent on these features. For example, the arc connecting root node X to node Y indicates that the target class X is dependent on of the feature Y and the P(X —Y) is higher than zero. Bayesian network has the Markov property if no directly connected nodes are conditionally dependent. The features within the Markov Blanket of the target class T node, MB(T), is the minimum set of features needed for classification [2, 27]. The Markov Blanket of a node A is the set of parent nodes PA and child nodes CA and the parent nodes of the child nodes PCA of node A in the Bayesian network. A wrapper classifier is applied further to drop non-required features without losing accuracy. The work in [13] proposes an optimal filter feature selection algorithm independent on any classifier for evaluating selected features based on an incremental association function assoc(). This function measures the degree/strength of association between each feature, CMB, to the target feature given the existing features in the CMB. The algorithm starts with an empty set CMB, adds from MB(T) the feature of maximum assoc() in the forward step then prunes the irrelevant features from CMB in the backward step. These steps are repeated forward and backward until the association of features to T vanishes given CMB.
- The problem of using the joint probability distribution function between a feature u and the target feature T with respect to another feature v is that it ignores the type of the correlation between the two features u and v. For example, if P(T, u, v) is greater than zero as shown in equation (2):

$$P(T , u, v) = P(T —u,v)*P(u—v)*P(v) \quad (2)$$

The case where features u and v are in the same dataset and P(T —u) > P(T —u,v) is ignored. This case can be clarified here in sample:

Consider a dataset containing ten instances, of two feature variables u and v and target class T as FTFTTFFFTT, TTTTTFFFFF, and FFFT- TFFTTT respectively.

Then:

P(T = true, u=true) = 0.8,
P(u= true, v=true) = 0.6,
P(T =true, u=true, v= true) = 0.66.

This leads to the observation that P (T=true, u=true) is greater than P (T=true, u=true, v=true), which means that considering feature u and feature v in a single dataset would decrease the discrimination of the target class T.

- Random forest is an ensemble technique based on building different decision trees from a random set of instances and a random subset of features. It considers the importance of each feature in the presence of a group of highly correlated features. The randomness in this algorithm ensures the variety of resulted decisions, and that the best split is applied according to the best result [15]. This technique permutes the values of each feature in the testing samples of each tree in the forest, then compare

the accuracy predicted before and after permuting in all trees. The rank of each feature increases as the average difference between the two predicted values over all trees increases. However, random forests technique may be prone to overfitting due the over emphasizing of noise in the input dataset and the performance's state of the art is not proved [12]. On the other hand, this technique selects the set of features randomly without considering the correlation among them.

- Evolutionary algorithms like swarm and ant colony algorithms are used for feature selection. Swarm colony algorithms describe the problem as a search space of states. The algorithm in [22] is based on the traditional forward feature selection strategy and hence suffer from the same problems described before. Genetic algorithm (GA) tries out large random populations of all possible sets of feature and assigns a fitness value to each depending on its performance on the classification task and until no further improvement is possible. Although the GA considers the interaction between features implicitly, there is no guarantee that the best combination of features is discovered [23].

3 Hypothesis Proof (Proof of Concept)

The stated hypothesis here is that every feature has either a positive, negative or neutral influence on the discriminating power of other features. If two features are included together in a subset to classify instances, the accuracy of classification could be enhanced, undermined, or not affected compared to using just one of these features to classify the instances. That is to say that some features can be considered complementary to each other while other features are considered incompatible to (contradicting with) each other. Features could also be neutral to each other. We hypothesis that utilizing this observation in a technique for feature subset selection would improve the discriminating power of the features selected.

Considering the forward feature selection technique, features are first ranked based on statistical techniques like Chi-Merge method. Ranked features are then added sequentially starting with the highest-ranked feature forming multiple feature subsets. Feature subsets are tested based on a classifier that is used as a fitness function for evaluating these features. The feature subset achieving the global maximum peak (the highest classification accuracy) is selected for further classification in the testing phase. Examining the curve of the classification accuracy of feature subsets, we observe that the curve does not go smoothly to a global maximum. Rather, it passes through a set of local maxima and local minima values as shown in figure 1. This behavior leads to the conclusion that some features have a constructive, destructive or neutral influence on each other. For example, feature two has a constructive effect on feature one, as including these two features in a single set shows an improvement in the classification accuracy resulting from using feature one only. Similarly, adding feature three has a constructive effect on the classification accuracy of using features one and two. Hence, we can conclude that the patterns within features one, two and three are considered complementary since the accuracy of classification is enhanced when the three features are

together in a feature subset. Features four and five have a neutral effect on the classification power of the feature subset containing features one, two and three. This is presumed due to the fact that the accuracy of classification shows no improvement after adding these features. Adding feature six to the feature subset set of features one to five deteriorates the classification accuracy to 50%. Indicating that feature six has a destructive effect on the discriminating power of the features from one to five, in spite of the fact that feature six on its own has a high correlation to the target class. The hypothesize here is that the values and hidden patterns within this feature contradict the patterns exploited by prior features, and hence feature six should not be included in the same feature subset as features one to five.

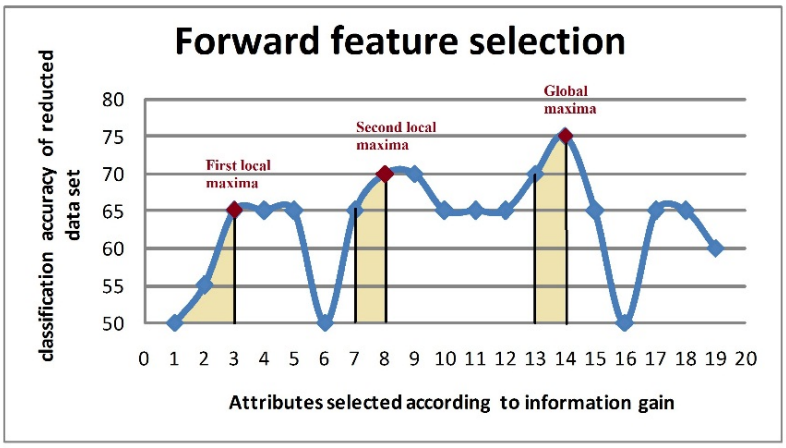


Fig. 1. Forward feature selection Technique behavior

The argue is about the usefulness of a forward feature selection technique that eliminates the features that cause local minima and hence resulting in a forward feature selection method that does not show local minima within its classification accuracy increasing curve.

Current selection techniques are classified into univariate methods that detect the patterns existing in the values of a single feature with respect to the target class, and multivariate methods that study the interconnection between features in addition to the correlation of each feature to the target class [24]. Information gain is a univariate technique that measures the mutual information between a single feature and the target class [26]. Mutual information MI (T; F) measures how much the uncertainty (entropy: marginal probability distribution) of a target class T is reduced if a feature F has been observed [25]. Measuring the relevance between features using information gain is biased towards features with higher value ranges [19, 20]. Chi-square and Chi-merge methods measure the dependency of an attribute on the target class labels according to the variance of the values of this attribute [5]. These univariate methods ignore the contribution of the multivariate patterns distributed among features in the classification of instances.

4 The Proposed Partnership-Gain Measure

The challenge that faces any feature selection technique is to find an optimal subset of features. According to the stated hypothesis, the optimal subset should include features that are complementary/constructive to each other. The proposed technique aims to search for the subset of features that maximizes the use of this hypothesis. The steps of the proposed technique are:

4.1 Step 1: Measuring interaction between features

This step aims to determine the value and direction of the interaction between pairs of features in the features set. First we measure λ_x which represents the percentage classification accuracy of a feature subset containing feature x and the target class. Next, we measure the classification accuracy of feature subsets of all possible pairs of features in the features set**:** λ_{xy} , λ_{xyz}, ... This step shows the nature and strength of the interaction (constructive, destructive or redundant) between pairs of features. Based on the stated hypothesis, the optimal subset should satisfy the *equation (3)* for every pair of features x and y in this subset:

$$\{\lambda xy > \lambda x\} \wedge \{\lambda xy > \lambda y\} \tag{3}$$

For example: Suppose a subset of 3 features x, y and z in a dataset, and a target feature t. The classification accuracy percentages *Acc* of x, y and z in separate feature subsets are 60%, 40% and 50% respectively. The classification accuracies of the feature subsets containing the pairs (x and y**)**, (y and z**)** and (z and x**)** are 70%, 40% and 40% respectively. Then the features x and y are complementary partners, while the features y and z are redundant partners, and the features x and z are undermining/contradicting partners to each other. Hence, a feature subset containing these three features is not an optimal subset of features to use for this classification task.

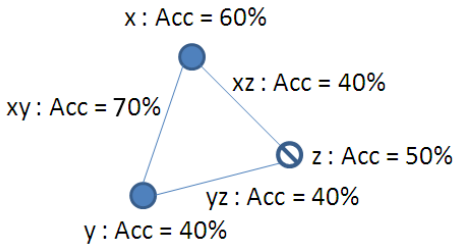


Fig. 2. The visualization of the interaction between three features in a dataset

4.2 Step 2: Calculating partnership-gain measure

Partnership-Gain PG_{xy} is calculated to be used in the proposed feature selection algorithm. Partnership-Gain PG_{xy} between features x and y is measured as shown in equation 4. Where E_{xy} is calculated according to equation 5. If the value of E_{xy} is greater than

λ_{xy} such that PG_{xy} has a negative value, then features x and y have a destructive effect on the performance of each other when included in the same feature subset. Otherwise, if the PG_{xy} has a positive value, then features x and y are complementary to each other.

$$PGxy = \lambda xy - Exy \tag{4}$$

$$Exy = (\lambda x + \lambda y)/2 \tag{5}$$

Equation 4 takes into consideration the special case where feature x has a greater contribution to the coupling of a pair (x, y) compared to the contribution of feature y as shown in equation 6.

$$\{\lambda_{xy} - \lambda_x\} >> \{\lambda_{xy} - \lambda_y\} \tag{6}$$

In some cases, four features x, y, a and b in a feature subset, λx and λ_y are greater than λ_a and λ_b while λ_{ab} is greater than λ_{xy}. In these cases, the importance of the single features like x and y should not be ignored in calculating the Partnership-Gain value PG_{xy}. The PG_{xy} is hence adjusted to be PG_{xy} as shown in equation 7, where a term that represents the maximum and minimum λ of all single features in the features set subset is added to the PG_{xy} to express the importance of features x and y with respect to the other features. The maximum and minimum λ among all single features is represented as max_λ and min_λ respectively.

$$\overline{PG_{xy}} = PG_{xy} + \frac{max_\lambda - min_\lambda}{max_\lambda - E_{xy}} \tag{7}$$

4.3 Step 3: Utilizing partnership-gain to select the optimal feature subset

This step selects the optimal subset of features based on the calculated PG of all possible pairs of features out of the n features in the features set. It searches for a combination of s features whose feature-pairs PG values are summed up and this sum is the maximum with respect to all the other combinations. Considering the PG values is an important factor for decreasing the complexity of computation rather than testing all the possible combinations. The sum PG_S of all Partnership-Gain PG_{xy} values for 0061ll pairs of features x and y in s is represented as follows in equation 8:

$$PG_s = \sum_{\forall x,y \in s} PG_{xy} \tag{8}$$

The maximum Partnership-Gain Measure value Max_{PG_S} is calculated as shown in equation 9:

$$Max_{PG_S} = max_{\forall s \in S}[PG_s] \tag{9}$$

The stated hypothesis here is that every feature has either a positive, negative or neutral influence on the discriminating power of other features. If two features are included together in a subset to classify instances, the accuracy of classification could be enhanced, undermined, or not affected compared to using just one of these features to

classify the instances. That is to say that some features can be considered complementary to each other while other features are considered incompatible to (contradicting with) each other. Features could also be neutral to each other. We hypothesis that utilizing this observation in a technique for feature subset selection would improve the discriminating power of the features selected.

5 The Proposed Feature Selection Algorithm

The proposed algorithm comprises two main parts. The first part calculates the ***PG*** values, and the second, the selection of the features based on these values is carried out. The input to the algorithm is the set of ***n*** features and the output is the optimal subset of features Max_S of size ***s***. The subset Max_S is selected based on the condition that it has the maximum sum value Max_{CG_S} of all partnership-gain values and the minimum size ***s***. An arbitrary filtering classifier is used in this algorithm like Naive Bayesian network classifier. The selection of the best combination of features starts by testing every feature ***x*** from the input dataset features. Then calculates the temporary value T_{sum} as the sum of the following values:

- $\sum_{y \in N \,\&\, y \neq x} PG_{xy}$: represents the sum of the ***PG*** values of every pair of features in the dataset *N* and the tested feature ***x*** in ***N***.
- $\sum_{Max_S} PG_{xy}$: represents the sum of the ***PG*** values of every pair of features in the selected subset of features Max_S and the tested feature ***x*** in ***N***.

The feature *x* that shows the maximum T_{sum} is selected to be added in the subset Max_S and removed from the input features set ***N***. This process is repeated in a while loop until the features set ***N*** contains no more features. Based on the forward feature selection method, the loop is repeated until TM_{sum} is not increasing anymore where the global maximum is reached or the dataset ***N*** contains no more features to add.

5.1 The feature selection algorithm: Partnership-gain algorithm to find the optimal feature subset for a classification problem

Input: Features set N of n features
Output: Subset Max_S: the optimal feature subset
//Calculating the Iambda values of all the single features in N
for $\forall\, x \in N$ **do**
$\lambda_x \leftarrow$ *the classification accuracy percentage of a feature set containing feature x*
end
//Calculating the PG values of all possible pairs of features in N
for $\forall$ *(x and y)* $\in N$ **do**
$\lambda_{xy} \leftarrow$ *the classification accuracy percentage of a feature set containing features x and y*
$E_{xy} \leftarrow (\lambda_x + \lambda_y)/2$
$PG_{xy} \leftarrow \lambda_{xy} - E_{xy}$

end
//Selecting the optimal subset of features
$t \leftarrow 0$: is a temporary variable
while $n>0$ **do**
sel is the selected feature
T_{sum} is a temporary sum of PG_s
$TM_{sum} \leftarrow 0$;
for $\forall\, x \in N$ **do**
$T_{sum} \leftarrow \lambda_x$
for $\forall\, y \in N$ *such that* $x = y$ **do** $T_{sum} = T_{sum} + PG_{xy}$ **end**
for $\forall\, y \in Max_S$ **do** $T_{sum} = T_{sum} + PG_{xy}$ **end**
if $T_{sum} > TM_{sum}$ **then**
$TM_{sum} \leftarrow T_{sum}$
$Sel \leftarrow x$
end
end
if $TM_{sum} > t$ **then**
set $n = n$ - 1;
set $t = TM_{sum}$;
Add feature *sel* to feature subset Max_S
Remove feature *sel* from the feature set N
end
else
Exit the loop and return Max_S
end
end
return Max_S

5.2 Algorithm complexity

To calculate the algorithm complexity, we will only focus on measuring the number of wrapper evaluations. In other words, we assume that measuring the classifying accuracy of instances in a dataset using a features subset to be O(**1**). Under this assumption, and if the number of features in the considered dataset is $\boldsymbol{n}$, then the total number of instructions needed to calculate the classification accuracy of all instances using one single feature is $\boldsymbol{n}$. The total number of instructions to calculate the accuracy of all instances in a dataset for all possible combinations of two features out of $\boldsymbol{n}$ is: $\frac{\boldsymbol{n!}}{\boldsymbol{2!*(n-2)!}}$.

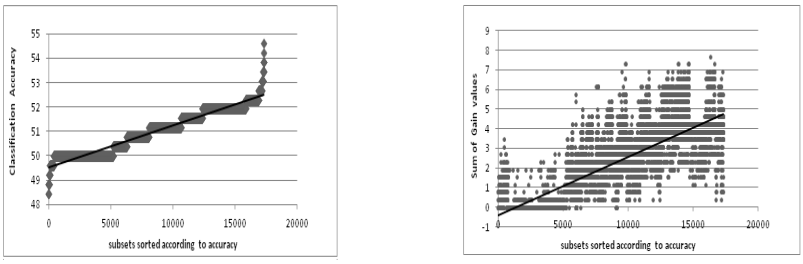
Accordingly, the complexity of the first part of the algorithm required to calculate the $\boldsymbol{\lambda}$ and $\boldsymbol{PG}$ values is of O($\boldsymbol{n^2}$). And the complexity of the second part of the algorithm required for selecting features is $\boldsymbol{n*(n-1)}$ which is simplified to O($\boldsymbol{n^2}$). Hence, the complexity of the whole algorithm is O($\boldsymbol{n^2}$).

6 Experimental Work

6.1 Partnership-gain measure proof of concept experiments

A set of prove of concept experiments are applied to clarify the importance of the Partnership-Gain Measure. The first experiment is applied on a benchmark Mutagenicity dataset cheminformatics and ChemAxon. This data set contains 23 extracted features and 260 instances divided equally into two categories to ensure the fairness of the experiment. These experiments tests the proportionality between the accuracy of classifying the data of a set of features and the Partnership-Gain Measure of these features. The classification accuracy test is the accuracy percentage of the 10-fold training and testing of the input dataset based on the Naive Bayesian tree method. In order to perform this test, a random number of features, f our features, will be chosen to be selected out of the 23 features. Although the basic Naive Byes algorithm assumes conditional independence between features, empirical results show that Naive Bayes works even if the independence assumption is ignored, Domingos and Pazzani (1997). The 10-fold cross validation experiment is applied on the data set to test the classification accuracy of a specific classifier after applying different feature selection techniques. Since the classifier is the same in all experiments, the resulted classification accuracy is dependent on the appropriateness of the selected features the tested feature selection technique. The number of combinations of 4 features out of the 23 features [C (23, 4)] is 17328 combination. Accordingly, the classification test will run 17328 time, and in each time the classification accuracy percent- age of a data set composed of the selected subset of four features and the Partnership-Gain Measure of this subset of features is recorded. Finally, a list of 17238 records are observed, these records are sorted based on the classification accuracy percentage. This sorted list is plotted once for classification accuracy and once for the corresponding Partnership-Gain Measure PG_S as shown in figure 3. The horizontal-X axis of the two charts in figure 3 represents the number of the combination (features subset) in the sorted list, such that the first point represent the features subset of the lowest classification accuracy. The vertical-Y axis of the first chart in this figure represents the classification accuracy while the vertical-Y axis of the second chart represents the Partnership-Gain Measure. The Partnership-Gain Measure of the selected features $\boldsymbol{PG_S}$ is the sum of the ($\boldsymbol{PG_{xy}}$) value of every pair of features x and y in the features set S is calculated as follows: $\sum_{\boldsymbol{y \in N \,\&\, y \neq x}} \boldsymbol{PG_{xy}}$

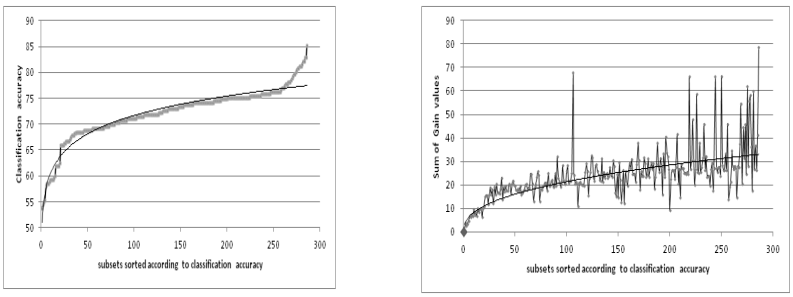
The trend line of both charts appears to be increasing with a similar slop, this shows that classification accuracy is directly proportional to the Partnership-Gain Measure. The features set Sm of the maximum Partnership-Gain Measure (Max ($\boldsymbol{PG_{xy}}$)) of this data set is considered as the selected set of features. The classification accuracy percentage of the data set that is composed of this selected subset of features Sm is 67.69%. When applying a classical feature selection technique like the Chi-Merge ranking method followed by forward selection, the data set of the selected subset of features by this technique shows as accuracy of 66.53%. This shows that the proposed method has higher classification accuracy percentage rather classical methods.



Accuracy vs combinations PG_S vs combinations

Fig. 3. Accuracy and PG_S vs combinations

The second experiment is applied on another benchmark dataset of heart data sound for illness detection. This data set is extracted from the UCI database and contains 270 instances and 13 features of two categories equally distributed. The same procedure of the first experiment is applied except that the tested combinations are of 3 features only out of the 13 features. The sorted list of the classification accuracy percentages and the Partnership- Gain Measures of the C (13, 3) combinations are plotted in the charts 4. Again the trend line of both charts proves that the classification accuracy is directly proportional to the Partnership-Gain Measure. On the other hand, the adjusted Partnership-Gain Measure ***PGs*** presented in equation 5 is utilized in this experiment. The scattering of the observed ***PGs*** values in the second chart in figure 4 is less than that of the observed ***PGs*** values in the second chart in figure 3 in the first experiment. This concludes the better accuracy of ***PGs*** rather than ***PGs***. The classification accuracy percentage of the data set that is composed of this selected subset of features S_m is 85.18%. The same result is achieved using Chi-Merge-Forward feature selection technique.



Accuracy vs combinations PG_S vs combinations

Fig. 4. Accuracy and $\overline{PG_S}$ vs combinations

6.2 Feature selection comparison to other techniques

Finally, for further proof of the competency of proposed ***algorithm 1***, the same comparison is applied on six dataset of various number of records and number of features.

The used datasets are Parkinsons, Hepatitis, Bupa, Ionosphere, Carcinogenicity and Mutagenicity, the number of features of each data set is 22, 19, 6, 34, 67 and 68 respectively, and the number of records of each data set is 96, 66, 288, 250, 292 and 4337 respectively. All of these data sets are categorized on to two class with equal number of records to grantee the fairness and accuracy in the extraction of rules. The source of the Parkinsons, Hepatitis, Bupa is the (UCI database 2007) from University of California in Irvine. And the sources of the Carcinogenicity and Mutagenicity datasets are the Benigni/Vari Carcinogenicity and the Bursi Mutagenicity benchmark databases in the cheminformatics website. The applied feature ranking techniques are ReliefF, ChiMerge, InfoGain and SVM techniques besides the proposed Partnership-Gain Measure. Then a forward feature selection techniques is applied based on the naive Bayesian tree classifier as an evaluation method. The feature selection and the used classifier are implemented in the Weka software from University of Waikato. The results of this experiment appears in *table 2* shows that the proposed ranking measure leads to a better results, either by increasing the classification accuracy or lowering the number of selected features, or both.

Table 1. Comparison analysis applied to different datasets.

Dataset	Used Technique	Classification accuracy %	Number of selected features
Parkinsons	Part-Gain	95.83	3
	ReliefF	94.79	8
	ChiMerge	93.75	2
	InfoGain	92.70	3
	SVM	92.70	11
Hebatitis	Part-Gain	68.18	3
	ReliefF	68.18	1
	ChiMerge	65.15	1
	InfoGain	65.15	1
	SVM	66.66	2
Buba	Part-Gain	68.18	3
	ReliefF	68.18	1
	ChiMerge	65.15	1
	InfoGain	65.15	1
	SVM	66.66	2
Ionosphere	Part-Gain	91.6	7
	ReliefF	91.2	7
	ChiMerge	89.2	2
	InfoGain	89.2	3
	SVM	90.0	13
Carcinogencit	Part-Gain	52.69	9
	ReliefF	52.69	16
	ChiMerge	51.53	3
	InfoGain	51.53	4
	SVM	51.53	15
Mutagenicity	Part-Gain	69.61	8
	ReliefF	66.15	11
	ChiMerge	66.53	4
	InfoGain	66.53	17
	SVM	66.15	8

7 Conclusion

This work proposes a new technique in the class of multivariate wrapper feature selection techniques. The proposed method is based a new measure: Partnership-Gain that calculates the degree of interdependency between features' subsets in addition to the correlation of individual features to each other and to the target feature. We argue that the partnership of features could in itself be constructive, destructive or neutral in terms of classification accuracy. The usefulness of this measure was proved practically on the classification task of various datasets. Results show direct proportionality between the classification accuracy and Partnership-Gain values of features' subsets of the datasets examined. In addition, results from various datasets show either an improvement or no change in the classification accuracy using the proposed technique. Examining the cases where no improvement in the accuracy was achieved re- veal that the features selected for inclusion in the feature subset were either not correlated or were all constructively correlated, hence the new technique produced the same feature set as the other techniques.

8 Reference

[1] Adair, J., Brownlee, A., & Ochoa, G. (2017). Evolutionary Algorithms with Linkage Information for Feature Selection in Brain Computer Interfaces. *In Advances in Computational Intelligence Systems* (pp. 287-307). Springer International Publishing. https://doi.org/10.1007/978-3-319-46562-3_19

[2] Aliferis, K. C. F., & Statnikov, A. (2014). U.S. Patent No. 8,655,821. Washington, DC: U.S. Patent and Trademark Office.

[3] Aliferis, C. F., Tsamardinos, I., & Statnikov, A. (2003). HITON: a novel Markov Blanket algorithm for optimal variable selection. In AMIA Annual Symposium Proceedings (Vol. 2003, p. 21). American Medical Informatics Association.

[4] Apolloni, J., Leguizamón, G., & Alba, E. (2016). Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. *Applied Soft Computing*, 38, 922-932. https://doi.org/10.1016/j.asoc.2015.10.037

[5] Bidgoli, A. M., & Parsa, M. N. (2012). A Hybrid Feature Selection by Resampling, Chi-squared and Consistency Evaluation Techniques. *World Academy of Science, Engineering and Technology*, 68, 276-285.

[6] Čehovin, L., & Bosnić, Z. (2010). Empirical evaluation of feature selection methods in classification. *Intelligent data analysis*, 14(3), 265-281. https://doi.org/10.3233/IDA-2010-0421

[7] Cao, D. S., Deng, Z. K., Zhu, M. F., Yao, Z. J., Dong, J., & Zhao, R. G. (2017). Ensemble partial least squares regression for descriptor selection, outlier detection, applicability domain assessment, and ensemble modeling in QSAR/QSPR modeling. *Journal of Chemometrics*, 31(11). https://doi.org/10.1002/cem.2922

[8] Cheminformatics, website : http://cheminformatics.org/datasets/

[9] ChemAxon Software, website : http://www.chemaxon.com/

[10] Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. *Artificial intelligence*, *151*(1-2), 155-176. https://doi.org/10.1016/S0004-3702(03)00079-1

[11] Dash, M., & Yee, O. C. (2007). extraRelief: improving relief by efficient selection of instances. *Lecture Notes in Computer Science*, *4830*, 305. https://doi.org/10.1007/978-3-540-76928-6_32

[12] Denil, M., Matheson, D., & De Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. *In International conference on machine learning* (ICML).

[13] Ditzler, G., Polikar, R., & Rosen, G. (2017). A Sequential Learning Approach for Scaling Up Filter-Based Feature Subset Selection. *IEEE Transactions on Neural Networks and Learning Systems*.

[14] Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, 29, 103-130. https://doi.org/10.1023/A:1007413511361

[15] Hapfelmeier, A., & Ulm, K. (2013). A new variable selection approach using random forests. *Computational Statistics & Data Analysis*, 60, 50-69. https://doi.org/10.1016/j.csda.2012.09.020

[16] Hua, J., Tembe, W. D., & Dougherty, E. R. (2009). Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42(3), 409-424. https://doi.org/10.1016/j.patcog.2008.08.001

[17] Lefakis, L., & Fleuret, F. (2016). Jointly informative feature selection made tractable by gaussian modeling. *Journal of Machine Learning Research*, 17(182), 1-39.

[18] Nguyen, M. H., & De la Torre, F. (2010). Optimal feature selection for support vector machines. *Pattern recognition*, 43(3), 584-591. https://doi.org/10.1016/j.patcog.2009.09.003

[19] Roy, A., Das, N., Saha, A., Sarkar, R., Basu, S., Kundu, M., & Nasipuri, M. (2015). A comparative study of feature ranking methods in recognition of handwritten numerals. *In Artificial Intelligence and Evolutionary Algorithms in Engineering Systems* (pp. 473-479). Springer, New Delhi. https://doi.org/10.1007/978-81-322-2126-5_52

[20] Saranyajothi, C., & D.thenmozhi, D. (2015). Machine learning approach to Document Classification using Concept based Features. *International Journal of Computer Applications*, 118(20), 33-36. https://doi.org/10.5120/20864-3578

[21] Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3), 297-336. https://doi.org/10.1023/A:1007614523901

[22] Schiezaro, M., & Pedrini, H. (2013). Data feature selection based on Artificial Bee Colony algorithm. *EURASIP Journal on Image and Video Processing*, 2013(1), 47. https://doi.org/10.1186/1687-5281-2013-47

[23] Siddiqui, M. A. (2016). An empirical evaluation of text classification and feature selection methods. *Artificial Intelligence Research*, 5(2). https://doi.org/10.5430/air.v5n2p70

[24] Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2016). A new hybrid filter–wrapper feature selection method for clustering based on ranking. *Neurocomputing*, 214, 866-880. https://doi.org/10.1016/j.neucom.2016.07.026

[25] Tourassi, G. D., Frederick, E. D., Markey, M. K., & Floyd, C. E. (2001). Application of the mutual information criterion for feature selection in computer-aided diagnosis. Medical Physics, 28(12), 2394-2402. https://doi.org/10.1118/1.1418724

[26] Yang, H., & Moody, J. (1999, June). Feature selection based on joint mutual information. *In Proceedings of international ICSC symposium on advances in intelligent data analysis* (pp. 22-25).

[27] Zeng, Y., Luo, J., & Lin, S. (2009, August). Classification using Markov blanket for feature selection. In Granular Computing, 2009, GRC'09. IEEE International Conference on (pp. 743-747). IEEE. https://doi.org/10.1109/GRC.2009.5255023

[28] Zhang, X. Y., Wang, S., Zhang, L., Zhang, C., & Li, C. (2016, April). Ensemble feature selection with discriminative and representative properties for malware detection. *In Computer Communications Workshops (INFOCOM WKSHPS), 2016 IEEE Conference on* (pp. 674-675). IEEE.
[29] University of California, Irvine (UCI) Machine Learning Repository, website: http://archive.ics.uci.edu/ml/
[30] University of Waikato (Weka) Data Mining Software in Java, website: http://www.cs.waikato.ac.nz/ml/weka/

9 Authors

Mostafa A. Salama completed his Doctorate in the field of Data Mining at Cairo University, faculty of Computers and Information in February 2012. He has several publications in peer reviewed international journals and conference proceedings in this field. His research interests are in applying and proposing Machine Learning techniques in the fields of Medical Informatics Analysis, Image Processing and Information Retrieval.

Ghada Hassan is a lecturer in the British University in Egypt since 2014. She taught the Programming in Java Course, Computer Science Department. She got here PhD in Computer Science in July 2010 from the University College London, UK in the area of Intelligent Systems in Finance.

Article submitted 09 November 2018. Resubmitted 23 December 2018. Final acceptance 15 January 2019. Final version published as submitted by the authors.