Determination of an Optimal Feature Selection Method Based on Maximum Shapley Value

Fatiha Mokdad[1,2], Djamel Bouchaffra[1], Nabil Zerrouki[1,2], Azzedine Touazi[1,2]
[1]Center for Development of Advanced Technologies,
Design and Implementation of Intelligent Machines Laboratory, Algeria
[2]University of Science and Technology Houari Boumediene, Algeria
E-mails: {fmokdad, dbouchaffra, nzerrouki, atouazi} @cdta.dz, f_mokdad@yahoo.fr

Abstract— **We propose a novel feature selection methodology based on game theory. In this context, the players are the various feature selection methods and the characteristic function (payoff) represents the feature ranking agreement within a coalition of players. The Shapley value assigned to each feature selection method is computed and ranked from higher to lower. The best feature selection method is identified as the one having the highest Shapley value. Finally, we have performed a score fusion scheme using the Borda Count (BC) consensus function as a benchmark to the maximum-Shapley value proposed approach. In order to validate the results obtained experimentally, we have performed a classification using a set of UCI and Statlog datasets by invoking an SVM classifier. Experimental results demonstrate the efficiency of the proposed methodology compared to some state-of-the-art approaches.**

Keywords- Feature selection; Shapley value; Borda Count consensus; SVM classification

I. INTRODUCTION

Automatic pattern classification remains an important task in various applications deployed nowadays. Before any classification task, one has to extract discriminative and less costly features. It is well known that the classifier power generalization depends essentially on the extracted features and the amount of data available (curse of dimensionality). Feature generation in pattern recognition represents a challenging and tedious task due to diverse data characteristics. The task of feature selection consists of disclosing a subset of uncorrelated and relevant variables from the initial set of available features without incurring much loss of information [1]. Feature selection has several potential benefits that aim at: (i) defying the curse of dimensionality to enhance the prediction performance, and (ii) reducing measurement and storage requirements as well as training and prediction times [2]. This paper focuses on the first issue, namely selecting input variables in an attempt to maximize the performance of a classifier on previously unseen data. It is worth noting that a large number of algorithms have been proposed in the literature for feature subset selection [3-8]. However, most of the proposed methodologies exhibit a common drawback: The selected features depend strongly on the application at hand, and there is no general feature selection scheme that provides an optimal classification with different data (e.g. nature and type of data) [3].

The main goal of this paper is to illustrate the value of exploiting game theory paradigm that attempts to quantify the interaction between players of a coalition and predict their optimal decisions. In our context, the mission consists of producing more powerful features that generates an optimal classification. The ranking agreement within a coalition of different feature selection models (players) is viewed as the payoff of a coalition. The players correspond to two recently proposed feature selection algorithms namely Fisher score [9] and Minimum Redundancy Maximum Relevance (mRMR) [2]. The mRMR method has been implemented with two feature evaluation metrics: They are the Mutual Information Difference (MID) and the Mutual Information Quotient (MIQ). We are finally considering three players within this game theory framework [2].

We believe that classification accuracy will increase when ones uncovers the feature selection method (or player) that contributes the most within a formed coalition. This contribution expressed by the maximum-Shapley value provides the optimal ranking of the features considered. This approach that relies on Shapley value has rarely been exploited in the feature selection literature. We have performed a score fusion scheme using the Borda Count (BC) consensus function as a benchmark to the maximum-Shapley value proposed methodology. Experiments conducted on the UCI and Statlog datasets for a classification task based on SVM show that the maximum-Shapley value feature selection, using BC fusion, can outperform or achieve similar classification results than other powerful feature selection methods.

The remainder of this paper is organized as follows. Section II introduces the contribution of coalitions using the maximum-Shapley value criterion, and the fusion strategy based on BC consensus function. In section III, we give a brief description of feature selection methods as well as the dataset used in the experiments. Section IV presents the evaluation measures. The obtained results, and discussion are detailed in section V. Section VI lays out the conclusion and future work.

II. PROPOSED METHODOLOGY

We first present the concept of Shapley value; afterwards, we describe the fusion idea based on BC consensus. The entire architecture of the proposed system is illustrated in Figure1.

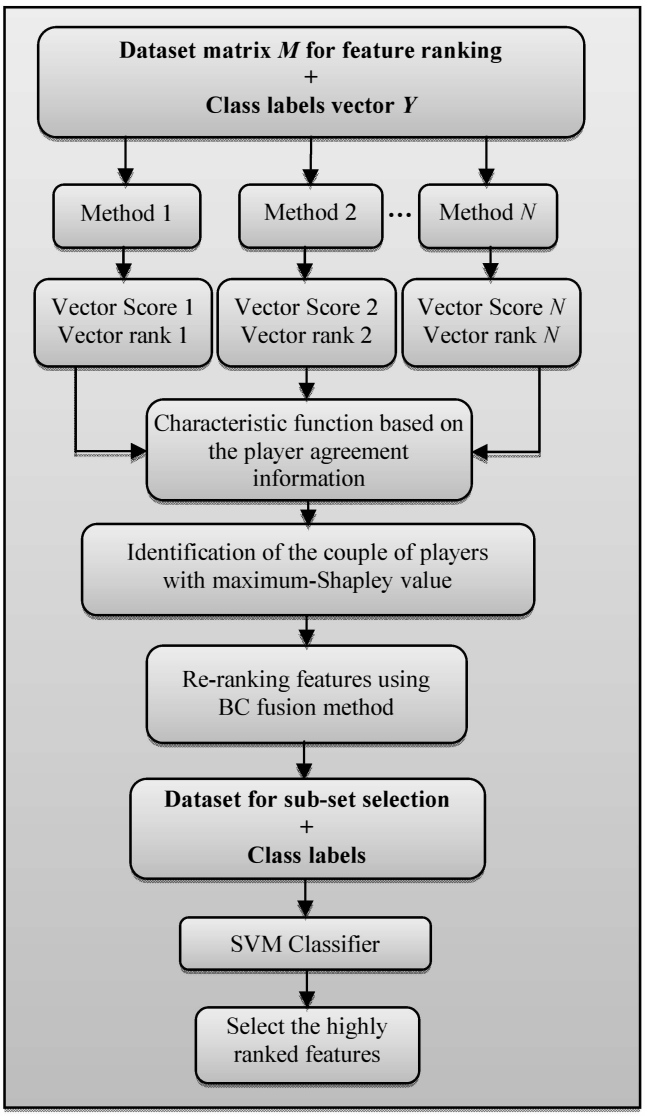


Figure 1. Flowchart of the maximum Shapley value-based feature selection methodology.

A. Proposed Shapley Value Approach

The main idea consists of determining how the feature selection methods are able to form a cooperative coalition with a maximum contribution [10]. Feature selection is performed from a given features dataset matrix $X=[x_{ij}]$ (i=1,…, m; j=1,…,n), with the corresponding class labels vector $Y=[y_j]$ (j=1,…,n). For each feature selection method, we have computed a score vector, sorted to produce the rank vector denoted as:

$$Sc=\begin{bmatrix}S_1 & S_2 & S_i & S_n\end{bmatrix} \quad , \quad R=\begin{bmatrix}R_1 & R_2 & R_i & R_n\end{bmatrix} \tag{1}$$

The next step consists of choosing the best coalition of two players among the N feature selection methods. This choice is based on the Shapley value criterion. First, the player's space is split into pairs of players without repetition and without permutation. Before obtaining the Shapley value for each player pair, one should calculate the characteristic function (V) (also known as payoff or utility function). This latter function is determined using the agreement information between the feature selection algorithms, which expresses how these methods can get together to form a coalition (C_i). The agreement, represented by $v(C_i)$, is measured by adding how many times the feature selectors agree on the ranking of a feature one by one. The resulting vector is denoted as:

$$V=\begin{bmatrix}v(C_1) & v(C_2) & v(C_3) & v(C_k)\end{bmatrix} \tag{2}$$

where k is the number of possible coalitions set to 2^M-1 and M=2.

In general, for a given (M, v) coalitional game *(where M players are considered as M feature selection methods)*, the Shapley value is defined as:

$$\Phi_i(v)=\sum_{C\subseteq M-i}\frac{|C|!(|M|-|C|-1)!}{|M|!}\{v(C\cup\{i\})-v(C)\}, i=1,\ldots,M \tag{3}$$

The Shapley values are then computed for each couple of feature selection method. Finally, the couple with the highest Shapley value is selected as the optimal one (the one that exhibits maximum payoff through cooperation).

The next step consists of fusing the two rank vectors corresponding to the highest Shapely value. To do that, we have invoked the traditional BC function, where a new feature ranking is computed.

B. Fusion Using Borda Count Consensus

The main idea behind the use of Borda Count is to re-rank the list of features in order of preference. This method has been introduced by Jean-Charles de Borda to resolve multi-classification problems, where each classifier is considered as a voter and classes are the candidates [11].

In this work, we used BC to determine the outcome of a feature selection by giving each feature a number of points corresponding to the number of rank (where each feature is a candidate and the two feature selection methods are the voters). Once all votes have been tallied the feature with the most points is declared the winner. Because it sometimes elects broadly acceptable candidates, rather than those preferred by a majority, the Borda Count is often described as a consensus based voting system rather than a majoritarian one. In fact, we have been used the BC method with the assumption that the re-ranking is based on points, thus the two feature selection algorithms (voters) are treated equitably [11].

III. EXPERIMENTAL FRAMEWORK

As previously mentioned, in this study we have implemented three feature selection methods namely: Fisher score and Minimum Redundancy Maximum Relevance (mRMR), with Mutual Information Difference and the Mutual Information Quotient. These selection methods are

considered as players, thus N=3, and the number of possible coalitions for each couple is M equal to 3. The Possible coalitions of feature selection methods denoted fs_i are: [{fs_1},{fs_2},{fs_1, fs_2}]. In this section, we briefly describe the three feature selection methods that are used in this work.

A. Selection methods discription

1) Fisher score: Fisher score is an affective filter-based linear feature selection method that assigns a weight to each feature based on Fisher criteria. Fisher score seeks features that exhibit a high discrimination power between classes [9]. It assigns the highest score to a feature that optimally separates the data. In other words, data points from different classes are far apart from each other whereas those pertaining to the same class remain nearby. Fisher criterion can also be used for feature extraction, such as Linear Discriminant Analysis (LDA). The major problem with Fisher score is that it processes one feature at a time, independently of the others. In other words, it does not consider combinations of features that might improve the performance [3].

2) Minimum Redundancy-Maximum Relevance (mRMR) for feature selection: The mRMR method selects the features based on their intrinsic characteristics. It determines their relevance or discriminant powers with regard to the targeted classes. The minimum redundancy criteria are supplemented by the usual maximum relevance criteria such as maximal mutual information with the target labels. It is expected that the selected feature set based on mRMR to be more representative of the target classes, therefore leading to a better generalization power. In other words, one can use a smaller mRMR feature set to effectively cover the same space as a larger feature set computed using conventional techniques. In this paper, we have decided to select the same number of features, ranked from higher to lower, according to their corresponding score values. To combine relevance and redundancy, we have used the selection criteria of a new feature: (1) MID: Mutual Information Difference criterion, and (2) MIQ: Mutual Information Quotient criterion. Please, do refer to Hanchuan Peng [2] for detailed information about the concept of mRMR.

B. Data collection

In order to evaluate the performance of the proposed feature selection method, we have conducted experiments on three databases obtained from the Statlog project and four databases obtained from the UCI Machine Learning Repository [12] (refer to Table I). “Satimag” dataset was generated from Landsat Multi-Spectral Scanner image data; it consists of the values of each pixel in a 3×3 neighborhood across 4 spectral bands, thus producing 36 features in total. “Splice” dataset consists of 3186 data points (splice junctions). The data points are described by 180 indicator binary variables and the problem is to recognize the 3 classes, i.e. the boundaries between exons and introns. “Abalone” dataset is predicting the age of the abalone; it contains 1 nominal and 7 continuous attributes with 4177 instances. The generated “Waveform” data set consists of 40 attributes with continuous values and a variable showing the 3 classes (33% for each of the three classes). Each class is generated from a combination of 2 of 3 “base” waves.

TABLE I. STATISTICS OF THE DATASETS USED IN THE EXPERIMENTS

Dataset	Training Data	Testing Data	Number of attributes	Classes
Satimag (Statlog)	4435	2000	36	6
Waveform (UCI)	3000	2000	40	3
Splice (UCI)	2000	1175	60	3
Abalone (UCI)	3133	1044	8	3
Vehicle Silhouettes (Statlog)	500	346	18	4
Heart (Statlog)	180	90	13	2
EEG Eye State (UCI)	8726	6249	14	2

The purpose of the “Vehicle” dataset is to classify a given silhouette as one of four types of vehicle. This latter may be viewed from one of many different angles. The features were extracted from the silhouettes by the Hierarchical Image Processing System (HIPS). “Heart” dataset contains 13 attributes with 270 observations where ‘1’ indicates the absence of heart disease and ‘0’ the presence of heart disease. The UCI data is acquired from one continuous EEG measurement with the Emotiv EEG Neuroheadset. The duration of the measurement was 117 seconds. The eye state (open or close) was detected via a camera during the EEG measurement and added later manually to the file after analyzing the video frames. '1' indicates the eye-closed and '0' the eye-open state.

C. Classification

In the experimental part, the proposed Maximum Shapley value methodology is implemented and tested on the training set of each dataset. In order to measure the efficiency of each feature selector method, we have invoked a Support Vector Machine (SVM) classifier trained with the Radial Basis Function (RBF) kernel. Furthermore, in order to exploit features of different types, and be able to optimally exploit the SVM, we have performed a scaling operation. Each feature domain has been mapped to the range [-1, +1].

In order to identify the optimal values of the cost and the kernel parameters (C, γ), respectively, we have applied a grid-search on C and γ to the training set [13]. Various pairs of (C, γ) values in which C= [2^{10}, 2^{9}, 2^{8} …, 2^{-5}] and γ= [2^{5}, 2^{4}, 2^{3} …, 2^{-10}] have been tested using fivefold cross-validation. Then, the best parameters set (C, γ) are chosen for accurately predicting the testing set.

IV. EVALUATION MEASURES

A. Overall accuracy

It is worth noting that the accuracy is the most widely used empirical measure [14]. However, accuracy does not discount the correct match between the reference and the predicted data obtained by pure chance. It is expressed as:

$$accuracy = \frac{tp + tn}{tp + fp + fn + tn}, \quad (4)$$

where *tp* is true positive, *tn* is true negative, *fp* is false positive, and *fn* is false negative.

However, to obtain an unbiased global accuracy, we have conducted a five-fold cross-validation procedure. The original sample for every class is randomly partitioned into five sub-samples. Out of these sub-samples, a single sub-sample is allotted for testing, and the remaining two sub-samples are saved for training. This process is repeated for five folds, with each of the five sub-samples used exactly once. Finally, a single overall value of the accuracy is computed through averaging.

B. Kappa coefficient

Another measure which can be extracted from a confusion matrix is the *Kappa coefficient*. It is a statistical measure of inter-raters agreement [14]. This measure is more robust than the simple accuracy measure since it subtracts the agreement occurring by chance. This coefficient is expressed as:

$$Kappa = \frac{P(a) - P(e)}{1 - P(e)}, \quad (5)$$

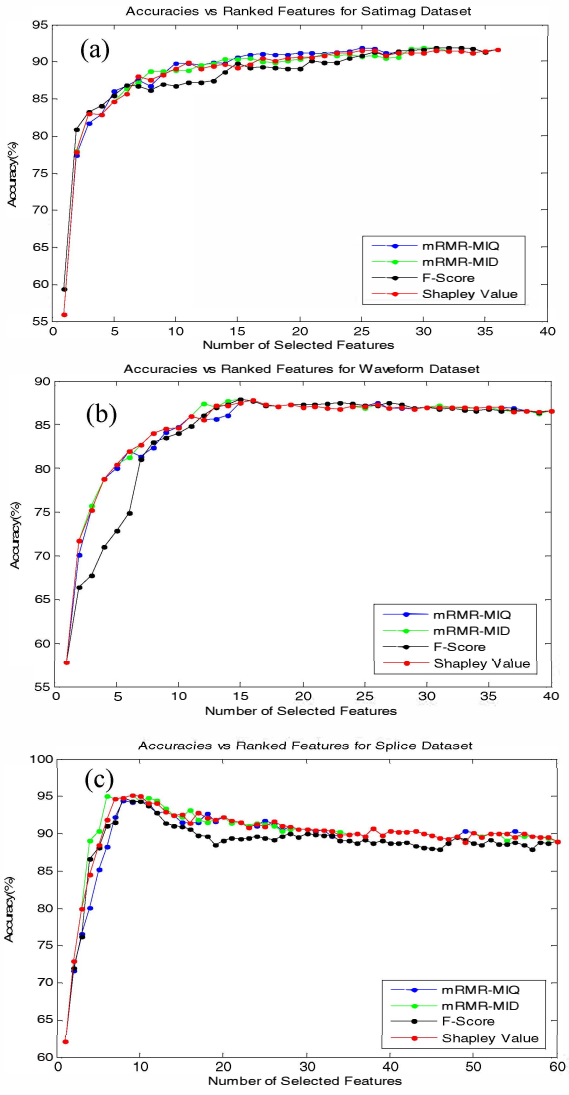
where $P(a)$ is the probability of relative observed agreement among raters and $P(e)$ is the probability of chance agreement. The range of the kappa coefficient is [-1, 1]. The value 1 represents perfect agreement, indicating that the raters agree in their classification in every case. The value 0 indicates agreement no better than chance, as if the rater has simply "guessed" every rating. A negative Kappa would indicate agreement worse than chance. In other words, a classifier exhibiting the highest value of the Kappa coefficients assigned to a set of classifiers is the most preferable.

V. RESULT AND DISCUSSION

We have compared the performance of the maximum Shapley-based feature selection method with other traditional feature selection techniques such as mRMRs (MID and MIQ) and fisher score. Two different types of evaluation measures such as the classification accuracy and the Cohen's Kappa coefficient were computed. Using the same principal as in frapper (filter-wrapper) types [15], the first feature ($f = 1$) is selected and the result of the classifier is assessed on the testing feature subset. If the accuracy of a given model after adding f's feature increases then the feature is added to the selected subset, and the next feature is tested until the increase is statistically significant.

A. Obtained results and interpretation

Two types of analysis have been conducted during this experimental phase: (i) comparison of the global classifier accuracy of the four feature selection methods with respect to the number of features selected (mRMRs, Fisher score, and the propped method). Fig. 2 shows the efficiency of the max-Shapley value-based feature selection method over the traditional ones, since its graph remains most of the time above the others namely, Waveform, Satimag and Splice databases. It is also indicated that this method is competing slightly with the mRMR-MID method. However after 25 features, one can observe that there is no significant difference between all four approaches. (ii) As reported in Table II, computation of the Kappa coefficient in a frapper framework and the accuracy for each feature selection method with and without BC fusion. The higher the Kappa value, the higher is the agreement between features to classify data. It leads to feature subset solutions with more relevant features.



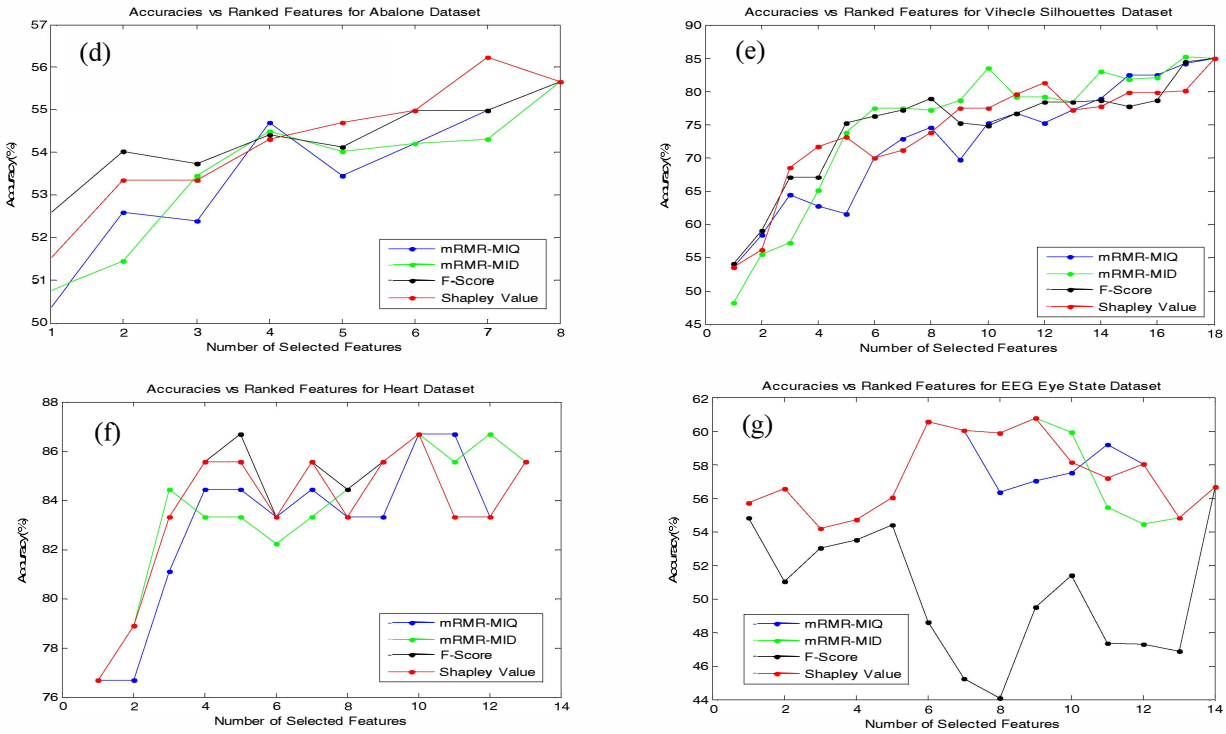


Figure 2. Comparaison of accuracy levels and number of features selected in the different datasets (a) Satimag dataset, (b) Waveform dataset, (c) Splice dataset, (d) Abalone dataset, (e) Vehicle Silhouettes dataset, (f) Heart dataset, and (g) EEG Eye State dataset.

TABLE II. STATISTICS OF THE DATASETS USED IN THE EXPERIMENTS

Algorithms / Databases	mRMR (MIQ)			mRMR (MID)			Fisher Score			Proposed BC-Shapley Value		
	(C, γ)	Accuracy	Kappa	(C, γ)	Accuracy	Kappa	(C, γ)	Accuracy	Kappa	(C, γ)	Accuracy	Kappa
Satimag	2^2 , 2^1	**91.95 (25)**	0.90	2^2-2^1	91.90 (31)	0.90	2^2-2^1	**91.95 (31)**	0.90	2^3-2^0	91.70 (36)	0.89
Waveform	2^1-2^{-4}	87.75(16)	0.81	2^2-2^{-6}	**87.95 (15)**	0.81	2^2-2^{-6}	**87.95(15)**	0.81	2^1-2^{-4}	87.75(16)	0.81
Splice	2^9-2^{-5}	94.46 (08)	0.91	2^9-2^{-6}	**95.14 (09)**	0.92	2^{10}-2^{-6}	94.80 (08)	0.91	2^9-2^{-6}	**95.14(09)**	0.92
Abalone	2^5-2^2	55.65(08)	0.33	2^5-2^2	55.65(08)	0.33	2^5-2^2	55.65(08)	0.33	2^2-2^4	**56.22(07)**	0.34
Vehicle Silhouettes	2^{10}-2^{-5}	84.97 (18)	0.83	2^{10}-2^{-5}	**85.26 (17)**	0.83	2^{10}-2^{-5}	84.97 (18)	0.83	2^{10}-2^{-5}	84.97(18)	0.83
Heart	2^3-2^{-7}	**86.66(10)**	0.72	2^3-2^{-7}	**86.66(10)**	0.72	2^{10}-2^{-9}	**86.66(04)**	0.72	2^3-2^{-7}	**86.66(10)**	0.72
EEG Eye State	2^0-2^5	60.55(06)	0.20	2^1-2^4	**60.77(09)**	0.16	2^3-2^3	56.68(14)	0.1	2^1-2^4	**60.77(09)**	0.16

VI. CONCLUSION AND FUTURE WORK

We have proposed a methodology that has the ability to determine the optimal feature selection methods based on the maximum Shapley value criterion. This coalitional approach seems to outperform traditional methods in some benchmarked databases. Because coalitional methods are not fusion methods, we have invoked the BC consensus function to investigate whether a fusion scheme is more appropriate than a coalitional scheme. Experimental results demonstrate the efficiency of the proposed BC-Shapely value compared to some relevant approaches. Future work consists of exploring other game theory tools such as the Banzhaf power index to identify the critical feature selection methods (players) within a coalition.

REFERENCES

[1] A. Jain, and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," Pattern Analysis and Machine Intelligence ,IEEE Transactions on, vol. 19, pp. 153-158, 1997.

[2] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE TPAMI, vol. 27, pp. 1226-1238, 2005.

[3] J. Kittler, M. Hatef, R.P. Duin, and J. Matas, "On Combining classifiers," IEEE TPAMI, vol. 20, pp. 226–239, 1998.

[4] C.F. Tsai, Y.C. Hsiao, "Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches," Decision Support Systems, vol. 50, pp. 258-269, 2010.

[5] K. E. Hild, D. Erdogmus, K.Torkkola, and J. C. Principe, "Feature extraction using information-theoretic learning," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, pp. 1385-1392, 2006.

[6] Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial intelligence, 97(1), 245-271.

[7] S. Cohen, G. Dror, and E. Ruppin, "Feature selection via coalitional game theory," Neur Comp, vol.19, pp.1939-1961, 2007.

[8] A. Touazi, F. Mokdad, and D. Bouchaffra, "Feature Selection Scheme Based on Zero-Sum Two-Player Game," 22nd International Conf. Pattern Recognition (ICPR), p. 1342-1347, 2014.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification. John Wiley & Sons, 2012.

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

[11] C. A. Perez, L. A. Cament, and L. E. Castillo, "Methodological improvement on local Gabor face recognition based on feature selection and enhanced Borda count," Pattern Recognition, vol. 44, pp. 951-963, 2011.

[12] UC Irvine Machine Learning Repository databases, University of California. URL: http://archive.ics.uci.edu/ml/.

[13] C. C. Chang and C. J. Lin, "LIBSVM : a library for support vector machines," ACM Transactions on Intelligent Systems and Technology, vol. 2, p. 27, 2011.

[14] W. Li and Q. Guo, "A new accuracy assessment method for one-class remote sensing classification," IEEE Trans. Geoscience & Remote Sensing, vol. 52, pp. 4621-4632, 2014.

[15] W. Duch, "Filter Methods. In: Feature extraction, foundations and applications," Studies in Fuzziness and Soft Computing, Eds. I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh, Physica-Verlag, Springer, 2006, pp. 89–118.