A NOVEL FEATURE SELECTION APPROACH APPLIED TO UNDERWATER OBJECT CLASSIFICATION

Tai Fei [1,2], Dieter Kraus [1] and Abdelhak M. Zoubir[2]

[1] IWSS, University of Applied Sciences Bremen, Neustadtswall 30, 28199 Bremen, Germany
[2] SPG at the Institute of Telecommunications, TU Darmstadt, Merckstraße 25,
64283 Darmstadt, Germany
{Tai.Fei, Dieter.Kraus}@hs-bremen.de, zoubir@spg.tu-darmstadt.de

ABSTRACT

A novel filter method for feature selection is presented. In our research, we observed that the feature relevance measures in the literature evaluate the features for classification purposes only with respect to certain aspects, e.g. distance, information theory, etc. Accordingly, the resulting feature selections may only be adapted to a narrow range of classifiers. Our approach jointly considers two relevance measures, i.e. mutual information (MI) and Relief weight (RW) so that the features are assessed more comprehensively. It requires not only the selection to hold sufficient MI, it also forces the features in the selection to have large RWs. In order to avoid an NP hard problem, a heuristic searching scheme is adopted, i.e. sequential forward searching. Moreover, the selection's cardinality can be determined automatically. Finally, this approach is applied to the underwater object classification and its classification results are compared to those of filter methods in the literature.

Index Terms— mutual information, Relief weight, filter method for feature selection, feature extraction, pattern recognition

1. INTRODUCTION

In the recent years, with the help of the modern synthetic aperture sonar (SAS) systems mounted on autonomous underwater vehicles (AUVs), the automatic target recognition (ATR) gains increasing attention. The ATR process is mainly composed of 3 steps: mine-like objects (MLOs) detection, feature extraction and mine type classification. First of all, a SAS image of a large region is scanned. The areas with suspicious objects are detected and marked. Then features describing the shape of objects and the textures of the sediments are extracted in these marked areas. Since there are a huge number of features available in the literature, we obtain a very large feature set. Due to the curse of dimensionality, the dimension of the space induced by the feature set should be reduced, e.g. dimensionality reduction, feature subset selection, etc. The principle component analysis (PCA) is a well known technique belonging to the dimensionality reduction. However, it is both sensitive to the data type and vulnerable to the scaling of the original data. Therefore, we prefer the feature subset selection, which draws a suitable subset out of the complete feature set. (e.g. filter and wrapper [1]). The problem of wrapper methods is that they highly depend on the chosen learning algorithm and are usually very time-consuming. Hence, filter methods should be favored. Instead of individual learning algorithms they use evaluation criteria such as mutual information (MI), Relief weight (RW) [2], etc. The MI is independent of the feature distribution and investigates the amount of classification relevant information contained in the features. It is widely adopted by the filter methods [3]-[5] like RELFSS, MIFS, and MISF-U. As for RELFSS, the MI of feature selections is normalized against their Shannon entropy (SE). A feature is selected according to the additional classification information that it can contribute. However, the normalization against SE implicitly incorporate criterion of minimum entropy. There is a risk of underfitting of the feature selection. MIFS and MIFS-U consider the sum of the MI of individual features and the redundancy between features is subtracted from the sum of MI. The problem is that the redundancy is not necessarily 100% classification relevant. Thus, the additional information which can be contributed by candidate features can be underestimated after removing the redundancy.

Furthermore, the dependency between selection and class index coded in terms of information entropy can be arbitrary. It is not always interpretable by classifiers. Therefore, we introduce a distance based measure, RW, in the feature selection process. Because of its efficiency a sequential forward searching (SFS) scheme is employed. Within every SFS cycle, there are two selection steps, i.e. RW selection and MI selection. Every selected feature should firstly pass the RW selection, which provides a set of features with relatively larger RWs. Then MI selection is ap-

This work was supported by ATLAS ELEKTRONIK Bremen and its research assistant Mr. Benjamin Lehmann. He helped us in calculating the features from the provided database.

plied to the set obtained in RW selection. Only the feature, which can contribute the largest MI to the feature selection in this set, will be considered as a useful feature and added to our selection. The evaluation is no longer constrained in individual aspects but the balance between them. The MI selection takes place after RW selection since the RW measures only the quality of individual features and provides nothing about the sufficiency of the selection. The MI does not only take care of the information contribution of incoming features but also the sufficiency of the selection. Hence, our approach can determine the cardinality of the feature selection automatically. This makes our feature selection process much faster than those filter methods which require manually setting the number of selected features. Finally, our approach is applied to the underwater object classification. Its classification results are then compared to those using MIFS, MIFS-U, RELFSS and mRMR [6].

2. FEATURE SELECTION ALGORITHM

Let $\boldsymbol{O} = \{X_1, X_2, \dots, X_N\}$ be the complete feature set of N features in total. A feature selection is a set denoted as $\boldsymbol{S} = \left\{X_{n_l} | n_l \in \{n_1, \dots, n_{N_s}\} \subseteq \{1, \dots, N\}\right\}$, where $N_s = |\boldsymbol{S}|$ is the number of selected features. A feature can be viewed as a random variable (RV). Therefore the feature value of a given instance in the database is a realization of the RV. Accordingly, $x_n^{(m)}$ is the m-th realization/instance of X_n, for $1 \leq n \leq N$ and $1 \leq m \leq M$. Furthermore, let C denote the class index, and $c^{(m)} \in \mathcal{C}$ is its m-th realization with $\mathcal{C}$ is the set of all possible class indices.

2.1. Relief Weight

The Relief algorithm given in [2] is a prominent filter selection method which evaluates individual features with a distance based relevance measure, RW. However, it was developed for binary-class problems. In this paper it is extended to the multiclass case. When feature X_n is taken into account, we find in the neighborhood of its m-th realization ($x_n^{(m)}$) 2 neighbors. One ($x_n^{(\text{hit})}$) is its nearest neighbor in the same class of $x_n^{(m)}$, and the other ($x_n^{(\text{mis})}$) is the nearest neighbor belonging to the classes which are different from the one of $x_n^{(m)}$. Employing the Euclidean distance, the weight assigned to the instance m is given as

$$w_n(m) = \left\| x_n^{(m)} - x_n^{(\text{mis})} \right\| - \left\| x_n^{(m)} - x_n^{(\text{hit})} \right\|. \quad (1)$$

Then the RW assigned to the feature X_n is

$$W_n = \sum_{m=1}^{M} w_n(m)\,. \quad (2)$$

The RW provides straight-forward information about whether the objects of different classes are overlapped or not in terms of the input feature. Accordingly, the larger the RW is, the better the feature is. Since the physical meaning of the individual features is various, their feature values can cover very different ranges, e.g. integers within the interval of $[0, 100]$, continuous probabilities between 0 and 1, etc. The resulting RWs could belong to different scales. The comparison between RWs of features is unfair. Thus, it is indispensible to convert the values of all the features into the same range. In this paper, all the features are scaled into the interval $[0,1]$ before RW computation.

2.2. Information Measure

The MI

$$I(X_n, C) = H(X_n) - H(X_n|C), \quad (3)$$

is a remarkable measure to investigate the classification relevant information contributed by features, where H denotes the Shannon entropy function. Moreover, the MI of a set of features is more important for us than the one of individual features given in (3), since the cardinality of $\boldsymbol{S}$, $N_{\boldsymbol{S}}$, is normally greater than 1. Therefore the joint mutual information (JMI) is required [7]. In the space $\mathbb{F}$ for $\dim(\mathbb{F}) = N_s$ induced by $\boldsymbol{S}$, an instance in the database is a point denoted by $\boldsymbol{x_s} = \left(x_1, \dots, x_{N_s}\right)^T$. Hence, the JMI between $\boldsymbol{S}$ and C can be defined as

$$I(\boldsymbol{S},C) = \sum_{\boldsymbol{x_s} \in \mathbb{F}} \sum_{c \in \mathcal{C}} p(\boldsymbol{x_s}, c) \log_2 \left(\frac{p(\boldsymbol{x_s}, c)}{p(\boldsymbol{x_s}) p(c)} \right). \quad (4)$$

We use the implementation given by Pocock in [8] to obtain the MI and JMI.

The removal of redundancy between individual features is an important issue in the methods such as MIFS, MIFS-U and mRMR. However, they all neglect the fact that the redundant information between features is not necessarily completely the classification relevant information. Moreover, complementary classification information can still be found among those features which possess high redundancy between each other [9]. In Fig. 1 although the redundancy is high, the additional information obtained from X_n is still significant. Therefore, what does matter is the amount of additional information contributed by feature X_n, which is quantified by

$$I_{\text{add}}(X_n) = I(\{\boldsymbol{S}, X_n\}, C) - I(\boldsymbol{S}, C). \quad (5)$$

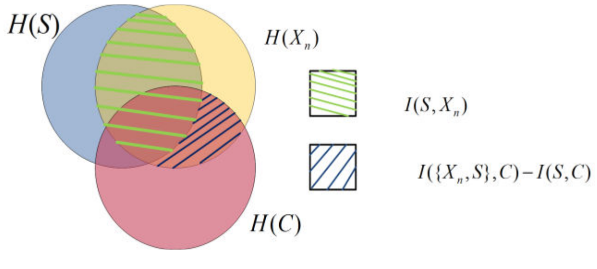

Fig. 1. Illustration of mutual information

2.3. Sequential Forward Searching With Combined Criterion

The complete searching space in our application is the set of all possible combinations of N_s features out of $N(1 \leq N_s \leq N)$. It causes an NP hard problem. We assume that our selection $\boldsymbol{S}$ is composed of only a small part of the complete feature set $\boldsymbol{O}$. Thus a SFS scheme is chosen to overcome this difficulty.

As already briefed in the introduction, there are two selection steps in every SFS cycle. In the RW selection, N_{cho} features, which possess larger RWs than the others, are picked up as candidate features from the set $\boldsymbol{O} \backslash \boldsymbol{S}$. In the MI selection, only the candidate, which maximizes the I_{add} in (5), is chosen to be added to the selection $\boldsymbol{S}$. The proposed algorithm called sequential forward searching scheme using Relief weight and mutual information (SFS-ReMu) is summarized as follows,

```
▪ begin, $\boldsymbol{S}$ is initialized as an empty set, and hence the remaining feature set $\boldsymbol{O}' = \boldsymbol{O}$. Let $n_{\mathrm{remain}} = |\boldsymbol{O}'|$.
    ○ do $n_{\mathrm{remain}} = n_{\mathrm{remain}} - 1$
        • calculate the $W_n$ of feature $X_n$, $\forall X_n \in \boldsymbol{O}'$,
        • find the $N_{\mathrm{cho}}$ features in $\boldsymbol{O}'$, which have the largest RWs, to compose a temporal set $\boldsymbol{S}_s$,
        • calculate the $I_{\mathrm{add}}(X_{n'})$, $\forall X_{n'} \in \boldsymbol{S}_s$
        • find the feature $X_k$, where
          $X_k = \arg\max_{X_{n'} \in \boldsymbol{S}_s} I_{\mathrm{add}}(X_{n'})$,
        • Add $X_k$ to $\boldsymbol{S}$, and $\boldsymbol{O}' \coloneqq \boldsymbol{O}' \backslash X_k$,
          if $|I(\boldsymbol{S},C) - I(\boldsymbol{O},C)| < \epsilon$, then break loop.
    ○ until $N_{\mathrm{cho}} > n_{\mathrm{remain}}$
▪ end
```

There is a free parameter N_{cho} in this approach. It controls the cardinality of the set $\boldsymbol{S}_s$, which contains the candidate features obtained in RW selection. If N_{cho} approaches M, SFS-ReMu behaves similarly as those methods which only maximize MI. On the contrary if it is close to 1, SFS-ReMu is similar to the Relief algorithm. We will discuss the choice of N_{cho} in the next section.

3. DATABASE AND NUMERICAL TESTS

3.1. Database Description

The database for testing the feature selection methods is provided by ATLAS ELEKTRONIK. There are in total 210 windows/instances, $M = 210$. Within every window there is one object: a truncated cone mine, a cylinder mine, or a stone as shown in Fig. 2.

The shape features in [10] are chosen to compose our feature set. Owing to the imperfectness of the contour extraction algorithms, the contours are smoothed before shape feature extraction. We also choose the mean value and the skewness of the power spectrum of the centroid distance of the object contours as shape features. In addition, the ring projection

$$f(r) = \int_0^{2\pi} u(r,\theta)d\theta \ , \qquad (6)$$

proposed in [11] is used, where $u(r,\theta)$ is a binary valued function in polar coordinates,

$$\begin{cases} u(r,\theta) = 1, \text{if } (r,\theta) \text{ locates within the contour} \\ u(r,\theta) = 0, \text{otherwise} \end{cases} . \qquad (7)$$

The skewness and weighted mean value of $f(r)$ are adopted as our shape features. The features of highlights and shadows are extracted separately, and those characterizing the relationship between highlights and shadows are included as well. Therefore we have totally 61 shape features. Furthermore, we take the co-occurrence matrix [12] and gray level run length matrix [13] to describe the texture. Due to the lack of *a priori* knowledge about parameter settings providing significant features, we allowed simultaneously several settings. Finally, we have 300 texture features. Thus there are totally 361 features, $N = 361$, in our feature set $\boldsymbol{O}$.

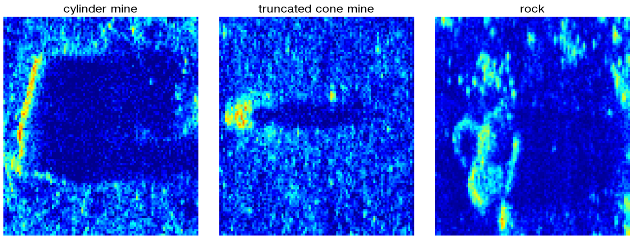

Fig. 2. Examples of objects.

3.2. Numerical Tests

First of all, the SFS-ReMu is applied to the database and we obtain several selections with different N_{cho}. Then classification tests are carried out with these selections. In the tests, five classifiers are used. PNN [14] is the probabilistic neural networks, KNN is the k-nearest neighbors, and KNN-DST [15] is the KNN assisted by Dempster-Shafer evidence theory. SVM-Gaussian and SVM-Poly denote the support vector machine (SVM) using a Gaussian and a polynomial kernel respectively. The classification rate ρ is defined as

$$\rho = \frac{m_{\mathrm{correct}}}{M} , \qquad (8)$$

where m_{correct} is the number of correctly classified instances. We use the leave-one-out scheme to make sure that every instance in the database is tested. In order to make the comparison among classifiers fair, every classifier is properly tuned so that it is able to achieve its best classification rate (ρ_{b}) by using the features provided by individual feature selection methods.

N_{cho}	1	2	3	4	5	6
N_s	8	11	10	9	8	8

Table 1. The number of selected features, SFS-ReMu.

N_{cho}	KNN	KNN-DST	PNN	SVM-Gaussian	SVM-Poly
1	0.7619	0.7714	0.7571	0.7952	0.7667
2	0.8095	0.8095	0.8095	0.819	0.7857
3	0.819	0.8048	0.7952	0.8238	0.7571
4	0.819	0.8048	0.7952	0.8238	0.781
5	0.819	0.8	0.7952	0.8333	0.8048
6	0.7905	0.7619	0.781	0.819	0.7857

Table 2. The best classification rates ($\boldsymbol{\rho}_\mathrm{b}$), SFS-ReMu.

In Table 1 the cardinality of the selection given by SFS-ReMu is listed. Only a small part of the features are chosen. It benefits the classification with less computation load and also avoids encountering the curse of dimensionality. In Table 2, the ρ_b of various classifiers using features selected by SFS-ReMu are recorded. The rows denote the results for the different N_cho settings, and the columns correspond to the classifiers. The best classification results are underlined in each column.

As discussed in the previous section, when $N_\mathrm{cho} = 1$, SFS-ReMu is very similar to the Relief algorithm, only those features with largest RWs are chosen. Its classification performance is accordingly not optimal. Moreover, when the N_cho increases, our method behaves more similar as MI based feature selection methods. As shown in the last row of Table 2, there is an obvious performance degradation when $N_\mathrm{cho} = 6$. Therefore, in the following discussion, only those results associated with $N_\mathrm{cho} = 2, 3, 4,$ and 5 are taken into account. When concentrated on the underlined results, we find that KNN-DST and PNN appreciate the setting of $N_\mathrm{cho} = 2$. It means that they are more inclined to choose the features with larger RWs. On the contrary, the features adapted to KNN, SVM-Gaussian and SVM-Poly should also contain rich MI in addition to large RWs. Thus it is preferable to include more features in the RW selection, and accordingly the N_cho is increased to 5.

Secondly, four algorithms known from the literature are implemented for comparison. Among them, the RELFSS selects the optimal features according to a MI based evaluation measure and it can also determine the cardinality of the selection automatically. The resulting N_S is 17. The ρ_b for the features obtained by RELFSS is listed in Table 3. However, compared with those in Table 2, RELFSS provides selections leading to a very poor classification performance.

	KNN	KNN-DST	PNN	SVM-Gaussian	SVM-Poly
ρ_b	0.6095	0.5905	0.6048	0.5905	0.3333

Table 3. The best classification rate ($\boldsymbol{\rho}_\mathrm{b}$), RELFSS.

KNN	KNN-DST	PNN	SVM-Gaussian	SVM-Poly
0.819(8)	0.8238(8)	0.8095(8)	0.8475(5)	0.8048(18)

Table 4. The best classification rates ($\boldsymbol{\rho}_\mathrm{b}$) and the number of features ($\boldsymbol{N_S}$), mRMR.

β	KNN	KNN-DST	PNN	SVM-Gaussian	SVM-Poly
0	0.7667(17)	0.7762(9)	0.7571(10)	0.8048(3)	0.7857(15)
0.3	0.7619(15)	0.8048(7)	0.7619(7)	0.8619(7)	0.7619(12)
0.5	0.7905(2)	0.7905(2)	0.7857(2)	0.8048(5)	0.8286(2)
0.7	0.7905(2)	0.7857(2)	0.7762(4)	0.7905(5)	0.8(10)
1	0.7238(1)	0.6333(3)	0.719(1)	0.7238(2)	0.6429(19)

Table 5. The best classification rates ($\boldsymbol{\rho}_\mathrm{b}$) and the number of features ($\boldsymbol{N_S}$), MIFS

β	KNN	KNN-DST	PNN	SVM-Gaussian	SVM-Poly
0	0.7667(17)	0.7762(9)	0.7571(10)	0.8048(3)	0.7857(15)
0.3	0.819(11)	0.8286(11)	0.8095(11)	0.8667(10)	0.8095(19)
0.5	0.7762(2)	0.7857(7)	0.7667(7)	0.8524(7)	0.8(8)
0.7	0.7905(6)	0.7857(6)	0.7619(6)	0.8476(6)	0.7905(10)
1	0.781(5)	0.7857(4)	0.7619(4)	0.8476(5)	0.7524(9)

Table 6. The best classification rates ($\boldsymbol{\rho}_\mathrm{b}$) and the number of features ($\boldsymbol{N_S}$), MIFS-U.

The classification results of mRMR, MIFS and MIFS-U are summarized in Table 4, Table 5 and Table 6 respectively. All of them require a manual setting of N_S. Peng *et al.* suggest in [6] trying a number of possible values of N_S and choose the one with the best classification rate. It is found in our numerical study that the selection cardinality N_S, which is larger than 20, can cause a dramatic performance degradation for the classification using our database. Accordingly, we try N_S from 1 to 20. Hence for every classifier, there are 20 candidate feature selections serving as inputs. All these candidates are then fed into the classifier. The candidate with the highest ρ_b is chosen. This ρ_b is recorded in the tables and so does its associated N_S in the brackets. Consequently, these three methods are very time-consuming due to the searching of optimal N_S across the 20 candidates. Obviously as shown in the tables, the optimal N_S is classifier-dependent. Hence, a fixed global setting of N_S for all the 5 classifiers would be improper.

Besides, the parameter β, with $0 \le \beta \le 1$, in Table 5 and Table 6 controls the tolerance of redundancy between features in MIFS and MIFS-U. When $\beta = 0$, the redundancy between features is completely ignored.

In summary, SFS-ReMu provides a fast feature selection procedure compared with mRMR, MIFS and MIFS-U. Its classification performance is also comparable to those of mRMR, MIFS and MIFS-U, and even outperforms them in most of the cases. Besides, PNN and KNN-DST prefer low value of N_cho, e.g. $N_\mathrm{cho} = 2$, while KNN, SVM-Gaussian and SVM-Poly favor higher value of N_cho such as $N_\mathrm{cho} = 5$.

4. CONCLUSION AND FEATURE WORK

The filter method for feature selection, SFS-ReMu, is presented. Compared with the existing methods, both MI and RW are considered in the feature relevance assessment steps. It provides us a fast feature selection procedure with an acceptable classification performance.

It is foreseen that we should try to build an evaluation function keeping the balance between mutual information and Relief weights so that the significance of individual measures in the joint consideration is adjustable. It is then possible to feasibly adapt the selection to different application.

5. REFERENCES

[1] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1, pp. 131-156, Mar. 1997.

[2] K. Kira, and L.A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," *Proceeding of AAAI-92*, pp.129-134, 1992.

[3] D.A. Bell and H. Wang, "A Formalism for Relevance and Its Application in Feature Subset Selection," *Machine Learning,* vol. 41, no. 2, pp. 175--195, 2000.

[4] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537-550, Jul. 1994.

[5] N. Kwak and C.H. Choi, "Input Feature Selection for Classification Problems,'' *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143-159, Jan. 2002.

[6] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no.8, pp. 1226-1238, Aug. 2005.

[7] H.H. Yang and J. Moody, "Feature Selection based on Joint Mutual Information," *Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis*, pp. 22-25, 1999.

[8] A. Pocock, *MITOOLBOX Matlab Toolbox*, University of Manchester, 2010.

[9] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, Mar. 2003.

[10] M. Yang, K. Kidiyo, and J. Ronsin "A survey of shape feature extraction techniques," *Pattern Recognition Techniques, Technology and Applications*. P.Y. Yin, Ed. Vukovar, Croatia: In-Teh, 2008, pp. 43-90.

[11] Y.Y. Tang, "Feature Extraction by Wavelet Sub-Patterns and Divider Dimensions," *Wavelet Theory Approach to Pattern Recognition*, 2nd edition, World Scientific Publishing Co. Pte. Ltd., Singapore, 2009.

[12] R.M. Haralick, K. Shanmugam, and I. Dinstein, "Texture Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610-621, Nov. 1973.

[13] J.S. Weszka, C.R. Dyer, and A. Rosenfeld, "A Comparative Study of Texture Measures for Terrain Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 4, pp. 269-285, Apr. 1976.

[14] D.F. Specht, "Probabilistic neural networks," *Neural Networks*, vol. 3, pp. 109-118, 1990.

[15] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 5, pp. 804-813, May 1995.