FEATURE SELECTION TECHNIQUES FOR CLASSIFICATION OF SATELLITE IMAGES WITH FUZZY ARTMAP NEURAL NETWORKS

DOMINIQUE RIVARD, ERIC GRANGER AND ROBERT SABOURIN
Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
École de technologie supérieure (ÉTS)
Montréal, Canada
rivard@livia.etsmtl.ca, {eric.granger, robert.sabourin}@etsmtl.ca

ABSTRACT
The selection of most relevant features can significantly reduce the resource requirement associated with fuzzy ARTMAP neural networks for the classification of high-dimensional data extracted from satellite imagery. This paper introduces a variant of a wrapper-type feature ranking technique that was previously proposed for ARTMAP neural networks by Parsons and Carpenter. As with the originally proposed technique, it evaluates the relevance of features based solely on between-class scatter from the geometry of internal ARTMAP categories. This paper also explores the inclusion into these feature ranking techniques of a within-class scatter measurement. Comparative simulations, performed on the Landsat multi-spectral imagery benchmark ('Satimage' from the StatLog repository), indicate that a significantly lower generalization error and fewer resources may be achieved by learning subsets produced by techniques that evaluate the relevance of features using both between- and within-class scatter.

I. INTRODUCTION

Satellite imagery from terrestrial observation conveys important information about land use and cover. For example, the classification of pixels in Landsat images can indicate the type of vegetation found in a landscape. The growing amount of data produced by spaceborne and airborne sensors, for a wide range of civilian and military applications, requires timely analysis. The high dimensionality of the data raises the curse of dimensionality in automatic pattern classification methods. To address this problem, the dimensionality of the feature space may be reduced using feature selection techniques.

Typically, feature selection algorithms consist of two components, a performance metric (to evaluate feature subsets) and a search engine (for finding those subsets.) Based on the performance metric, Langley (1994) defines two types of approaches to feature selection. When the feature selection algorithm evaluates feature subsets based solely on the data, regardless of the classifier for which it is intended, the algorithm is a *filter* method. The problem with filter methods is that the selected feature subset can be inappropriate for the classifier in mind. In contrast, a *wrapper* method exploits the classifier in its performance metric. Regardless of the performance metric, the search engine is faced with the problem of finding the best candidate feature subsets in an exponential solution space. For a problem with m features, the search space consists of 2^m solutions. Finding the optimal solution demands an exhaustive search, which is impractical even for data with moderate dimensionality. For this reason, sub-optimal

search strategies are typically employed. One such strategy is called feature ranking, a relaxed form of feature selection that ranks features with respect to their relevance.

The fuzzy ARTMAP (FAM) neural network architecture, proposed by Carpenter *et al.* (1992) is capable of self-organizing stable recognition categories in response to arbitrary sequences of analog and binary input patterns. It provides a unique solution to the stability-plasticity dilemma faced by autonomous learning systems. Since FAM can perform fast, stable, on-line, unsupervised or supervised, incremental learning, it can learn from novel events encountered in the field, yet overcome the problem of catastrophic forgetting associated with many popular neural networks classifiers. It has been successfully applied in complex real-world pattern recognition tasks such as the recognition of radar signals (Granger *et al.*, 2000), multi-sensor image fusion (Carpenter *et al.*, 2004), remote sensing and data mining (Mamman *et al.*, 1998), handwriting recognition (Bote-Lorenzo *et al.* 2003), and signature verification (Murshed *et al.*, 1997).

Streilein *et al.* (2000) have presented a feature selection wrapper-type method that is specialized for the FAM neural network. It considers the amount of overlap in feature space between target and non-target classes, encoded in the internal category nodes at the F2 layer of the FAM, in order to rank features in the input feature vector. Therefore, it exploits between-class scatter to rank features. Parsons and Carpenter (2003) apply this technique to other neural networks in the ARTMAP family (default ARTMAP, ARTMAP-IC and distributed ARTMAP), and extend it to allow for on-line learning of an arbitrary number of target classes.

In this paper, a new within-class scatter measurement is combined with this between-class measure to improve the quality of feature subsets. In addition, another between-class scatter measure that is more consistent with this proposed within-class measurement is introduced. This paper is organized as follow. In the next section, FAM is briefly reviewed and an overview of the existing wrapper-type algorithm for feature selection adapted to ARTMAP classifiers is presented. Then, between- and within-class scatter measures are discussed and new metrics for feature ranking are proposed. Section IV describes the experimental methodology and the Landsat dataset used for proof of concept simulations. Finally, simulation results are presented and discussed in Section V.

II. FEATURE SELECTION FOR FUZZY ARTMAP NEURAL NETWORKS

The FAM neural network consists of two fully connected layers of nodes: an M node input layer, F_1, and an N node competitive layer, F_2. A set of real-valued weights $\mathbf{W} = \{w_{ij} \in [0,1]: i = 1, 2, \ldots, M; j = 1, 2, \ldots, N\}$ is associated with the F_1-to-F_2 layer connections. If complement coding is used, $M = 2m$, where m is the dimensionality of input patterns. Each F_2 node j represents a recognition category that learns a prototype vector $\mathbf{w}_j = (w_{1j}, w_{2j}, \ldots, w_{Mj})$. The F_2 layer of FAM is connected, through learned associative links, to an L node map field F^{ab} , where L is the number of classes in the output space. A set of binary weights $\mathbf{W}^{ab} = \{ w_{jk}^{ab} \in \{0,1\}: j = 1, 2, \ldots, N; k = 1, 2, \ldots, L\}$ is associated with the F_2-to-F^{ab} connections. The vector $\mathbf{w}_j^{ab} = \left(w_{j1}^{ab}, w_{j2}^{ab}, \ldots, w_{jL}^{ab}\right)$ links F_2 node j to one of the L output classes.

After an ARTMAP has been trained on complete input vectors of the training set, the feature selection wrapper-based method proposed by Parsons and Carpenter (2003) first ranks the features by relevance. With this wrapper, the relevance of features is based on the degree of overlap of the weight intervals as calculated by:

$$D(k \mid i) = \sum_{\substack{J:W_{Jk}=1 \\ j:W_{jk}=0}} \left[\frac{\left| w_{i+m,J} - w_{i+m,j} \right| + \left| w_{iJ} - w_{ij} \right|}{1 - \left(w_{i+m,J} \wedge w_{i+m,j} \right) - \left(w_{iJ} \wedge w_{ij} \right)} \right] \tag{1}$$

Where $\wedge$ is the fuzzy AND minimum operator, J represents the indices of all weights associated to class k ($J : W_{Jk} = 1$) and j represents the indices of all weights associated to classes other than k ($j : W_{jk} = 0$). $D(k|i)$ indicates how well the feature i alone differentiates the output class k from all the other classes. Interpretation of Eq. (1) is straightforward in the case of two categories. The numerator measures the degree of non-overlap between two ARTMAP categories, while the denominator measures the length of the interval covered by the two categories. $D(k|i)$ results in a value of 0 if the two categories coincide perfectly, a value between 0 and 1 if the categories overlap, a value equal to 1 if the categories are juxtaposed, and a value between 1 and 2 if an interval separates the two categories. If there are more ARTMAP categories, the two embedded summations makes the result of $D(k|i)$ harder to interpret directly.

For each class k, the features are sorted by decreasing value of $D(k|i)$. Let $O(k|i)$ be the position of the feature i in the ordered list. The relevance of feature i is equal to:

$$D(i) = \sum_{k} \frac{1}{O(k \mid i)} \tag{2}$$

Then, features are ranked according to their $D(i)$ value. Once the feature ranking is known, a variant of the sequential forward search (SFS) (see Jain *et al.*, 2000 for a description of the original algorithm) oriented by the ranking is performed. This algorithm can be described as follow:

1. Start with the empty feature subset $\Phi = \{\varnothing\}$
2. Select the feature i with the next largest $D(i)$
3. Train the neural network and test on a validation set using the feature subset $\Phi \cup i$; if the classifier performance improves, update $\Phi \leftarrow \Phi \cup i$
4. Go to step 2

III. BETWEEN- AND WITHIN-CLASS MEASURES FOR FEATURE RANKING

Linear Discriminant Analysis, a supervised feature extraction method proposed by Fisher (1936), uses between- and within-class scatter measures to evaluate the quality of input features. However, the feature selection method presented in Section II is based solely on between-class measurement of scatter.

In an ARTMAP memory representation of the feature space, between-class scatter corresponds to the overlap between categories of different classes. In contrast, within-class scatter corresponds to the spread interval covered by the categories associated to a given class. A greater between-class scatter and a smaller within-class scatter should provide stronger insight into the quality of input features.

A within-class measurement adapted to ARTMAP neural networks should indicate the compactness of the categories associated with a given class. Let $w_i^{\min} = \min\{w_i : i = 1, 2, \ldots, N\}$ be the smallest value of ARTMAP weight for the dimension i. Then, the compactness may be defined as:

$$C(k \mid i) = 1 - w_{iJ}^{\min} - w_{i+m,J}^{\min} \tag{3}$$

where $C(k|i) \in [0,1]$ is the spread ratio according to dimension i of the categories associated to class k.

Between- and within-class scatter may be combined into one metric, by considering that the quality of an input feature is proportional to the first, and inversely proportional to the later. The between-class scatter measure of Eq. (1) may be combined with the within-class scatter measure of Eq. (3) in the form:

$$D^{\text{bwc}}\left(k \mid i\right) = \frac{D\left(k \mid i\right)}{C\left(k \mid i\right)} \tag{4}$$

A between-class scatter measure that is more consistent with Eq. (3) is:

$$D_1\left(k \mid i\right) = \frac{\left|w_{i+m,J}^{\min} - w_{i+m,j}^{\min}\right| + \left|w_{iJ}^{\min} - w_{ij}^{\min}\right|}{1 - \left(w_{i+m,J}^{\min} \wedge w_{i+m,j}^{\min}\right) - \left(w_{iJ}^{\min} \wedge w_{ij}^{\min}\right)} \tag{5}$$

where $D_1(k|i) \in [0,2]$. This variant of between-class scatter is closely related to the measure proposed by Streilein *et al.* (2000). However, it follows the same principle of measurement as $C(k|i)$. It focuses on the ranges of categories, and neglects the number of categories per class. This may cause the measure to be less discriminant when the intervals covered by the categories of each class span similar ranges of values. Although the interpretation of $D_1(k|i)$ is identical to that of $D(k|i)$, the final result is always directly interpretable because there are no summations in the equation. For this same reason, it has a lower algorithmic complexity, even though all ARTMAP categories must be compared to find the smallest value.

IV. EXPERIMENTAL METHODOLOGY

For proof of concept, comparative simulations have been conducted on Landsat multispectral imagery. The data set employed for simulations is the 'Satimage' benchmark used originally by King *et al.* (1995) and Michie *et al.* (1994) in the StatLog project, and is available at the UCI Repository of machine learning databases[1]. The data set consists of 6435 samples with 36 features, belonging to one of six classes. The samples come from an 82×100 pixels sub-area of a scene taken by the Landsat Multispectral Scanner (MSS) with four bands (green, red and two near infrared). Each pixel has an 80 meters resolution cell and was labeled on site by an analyst. An input pattern corresponding to a pixel is formed by concatenating the four features (on per spectral band) of that pixel with the 32 features of its 8 neighbors. Note that the labels are assigned based on the original four features (f_{17} to f_{20}) of each pixel.

The dataset is partitioned into k-folds for cross-validation with $k = 10$. For each experiment, one fold is used as a hold-out data set and the nine others are partitioned into three data sets (training, validation and selection) containing respectively 50%, 25% and 25% of the remaining examples. The priors of each class are respected through all data sets. The hold-out data set is normalized using the min-max technique with the parameters estimated on the training, validation and selection data sets to avoid the introduction of an experimental bias.

For each experiment, a FAM is constructed using the training data set, through hold-out validation (using the validation data set) to avoid overfitting. Fuzzy ARTMAP

[1] http://www.ics.uci.edu/~mlearn/MLRepository.html

parameter values are fixed at α = 0.001, β = 1, $\bar{\rho} = 0$ and ε = -0.001. To reduce the intrinsic variance of FAM, the training is repeated 10 times, and the selection data set is used to select the best of 10 classifiers. The feature selection wrapper method, using the different between-class and within-class scatter measurements, is applied to the resulting FAM. In Step 3 of the modified SFS algorithm, training and validation data sets are used to train the neural network, while selection data set is used to evaluate its performance. From the ten resulting feature subsets, a final feature selection is performed to reduce the variance of the feature selection method. Only the features present in at least T of these subsets are kept in the final feature subset. The threshold value T may be adjusted to increase or decrease the size of the final feature subset. For the current experimental methodology, T = 5. Ten new neural networks are constructed with the final feature subset using the same methodology to evaluate the generalization error on the hold-out data set.

The performance of feature ranking techniques, based on between- and within-class scatter metrics, are compared based on generalization error, size of the final feature subset and compression ratio. Compression ratio represents the mean number of patterns per ARTMAP categories. Generalization error represents the ratio of misclassified patterns on the hold-out data set.

V. RESULTS AND DISCUSSION

The histograms shown in Figure 1 depict the subsets obtained with the feature selection wrapper methods for the 10 folds. The higher counts of Figure 1b indicate that the inclusion of within-class scatter in the metrics provides more stability in feature selection.

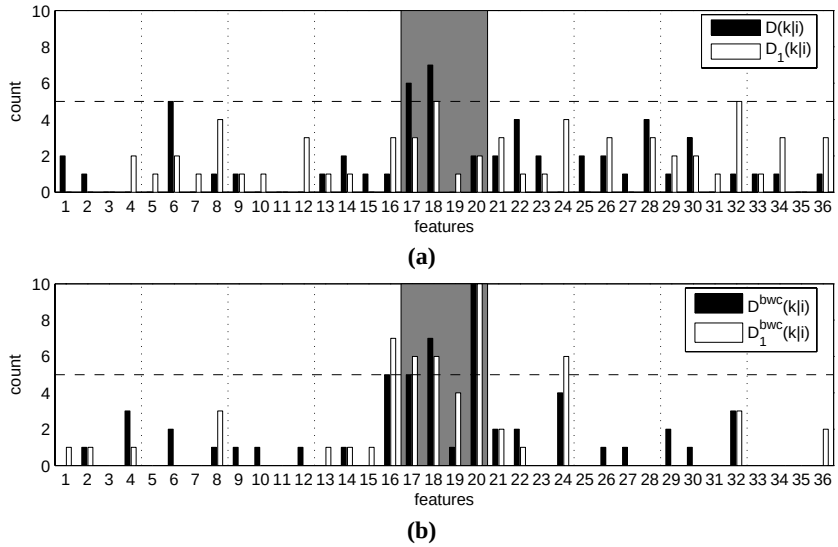

Figure 1 Features selected for (a) between-class only and (b) between- and within-class feature ranking metrics. The 36 features are divided into 9 groups of 4 features, one group per pixel. The area in gray corresponds to the central pixel.

Table 1 shows the final feature subsets for each feature ranking metrics. Metrics using both between- and within-class scatter measure selected three (f_{17}, f_{18} and f_{20}) of the four features belonging to the central pixel. Recall that features f_{17} to f_{20} correspond to the features measured for the center pixel (shown in bold).

Table 1 Final selected feature subsets for each feature ranking metrics.

Metric	Subset	Size	Metric	Subset	Size
$D(k\|i)$	$\{f_6, \mathbf{f_{17}}, \mathbf{f_{18}}\}$	3	$D^{\text{bwc}}(k\|i)$	$\{f_{16}, \mathbf{f_{17}}, \mathbf{f_{18}}, \mathbf{f_{20}}\}$	4
$D_1(k\|i)$	$\{\mathbf{f_{18}}, f_{32}\}$	2	$D_1^{\text{bwc}}(k\|i)$	$\{f_{16}, \mathbf{f_{17}}, \mathbf{f_{18}}, \mathbf{f_{20}}, f_{24}\}$	5

Figure 2 shows the generalization error obtained on the hold-out data sets. The box and whisker plots are presented in ascending order of median. Boxes whose notches do not overlap indicate that the median of the two groups differ at the 5% significance level. The results obtained using the complete feature set (label FAM R36), and the results obtained using only the four features of each pixel (label FAM R4) are presented for reference. Lowest generalization error is obtained from a nearest neighbor classifier using the complete feature set. This last result confirms those obtained by King *et al.* (1995).

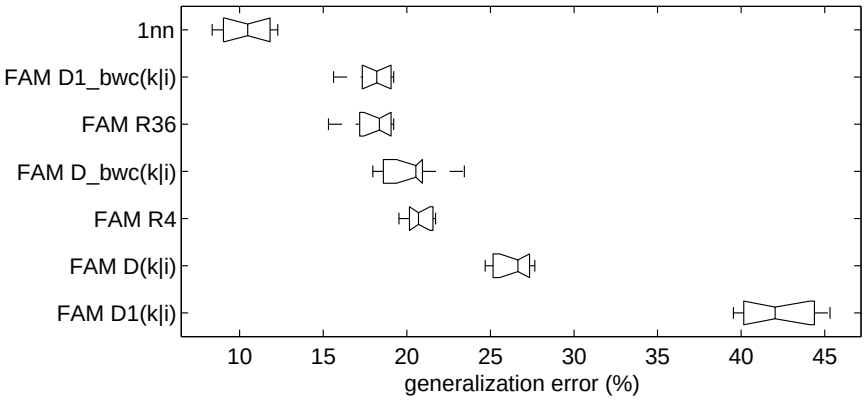

Figure 2 Generalization error on the hold-out datasets.

Generalization error obtained using only the four features of each central pixel is significantly higher than the one using the complete feature set. This is caused by *inconsistent cases* in the 'Satimage' dataset. When considering only the features of each pixel, 1014 patterns (15.8% of the samples) have identical feature vectors with different class labels. These inconsistent cases augment the amount of overlap between classes and the effectiveness of feature ranking metrics is reduced. $D_1(k|i)$ is especially vulnerable because it measures the interval covered by all the categories, regardless of the number of categories or their location. Hence, the final feature subset it generates is poor. The inconsistent cases are "eliminated" by adding the 32 features from the 8 neighboring pixels.

The generalization error from final feature subsets obtained with metrics combining between- and within-class scatter measure are significantly lower than their counterparts using only between-class scatter measures. The generalization error obtained from the

final feature subset selected using $D^{\text{bwc}}(k \mid i)$ offers no significant difference from the four features of each central pixel. As such, it is significantly higher then the generalization error obtained from the final feature subset selected using $D_1^{\text{bwc}}(k \mid i)$, which offers no significant difference from the complete feature set. $D_1(k|i)$ is the metric that benefits the most from the inclusion of the within-class scatter measure.

Figure 3 shows the generalization error as a function of the ARTMAP compression. Neural networks constructed with the complete feature set show the best compression. This is a direct effect of the dimensionality of the data as the number of features is increased. Without increasing the number of patterns, the feature space is more sparsely populated. The ARTMAP categories tend to overlap less, hence fewer categories are needed to map the feature space. However, neural networks produced as a result of the feature selection methods use fewer features, which compensates their lower compression ratios. This result shows that compression provides less insight in evaluating feature selection wrappers for the FAM than generalization error and size of feature subsets. However, Figure 3 indicates the possible trade-off between generalization error and feature subset size.

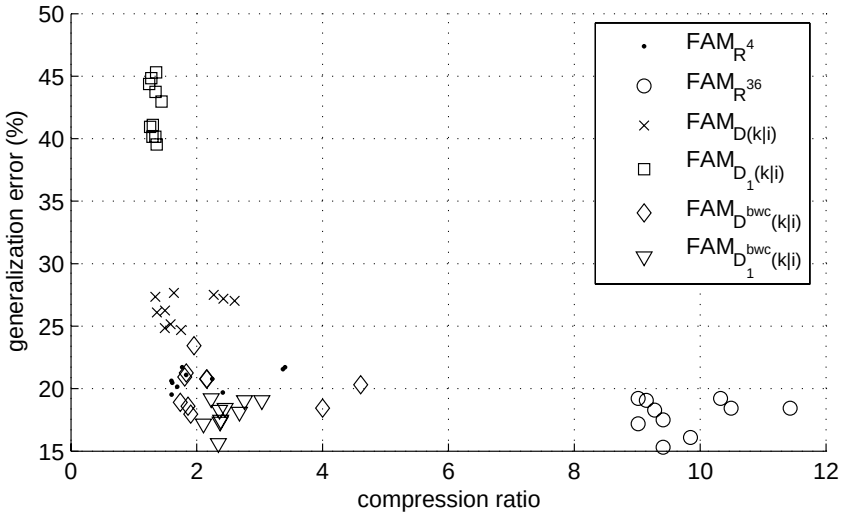

Figure 3 Generalization error in function of ARTMAP compression.

VI. CONCLUSIONS

In this paper, feature selection adapted to FAM has been explored in the context of classification of pixels extracted from satellite imagery. A between-class scatter based feature selection wrapper algorithm proposed by Parsons and Carpenter (2003) and adapted to ARTMAP neural networks has been reviewed. A new within-class scatter measurement has been combined with this between-class scatter measure to improve the quality of feature subsets. In addition, another between-class scatter measure that is more consistent with this proposed within-class measurement has been introduced. These

measures have all been compared by proof of concept simulations performed on the Landsat multispectral imagery benchmark ('Satimage' from the StatLog repository).

The results indicate less variance in feature selection methods using both between- and within-class scatter measures. In addition, feature subsets produced by these methods have a significantly lower generalization error on hold-out data, compared to methods using only a between-class scatter measure. Finally, the results show the presence of a trade-off between generalization error and feature subset size.

ACKNOWLEDGEMENTS

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

Bote-Lorenzo, M. L., Dimitriadis, Y. A., and Gomez-Sanchez, E., 2003, "Automatic Extraction of Human-Recognizable Shape and Execution Prototypes of Handwritten Characters," *Pattern Recognition*, Vol. 36, pp. 1605-1617.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., and Rosen, D. B., 1992, "Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps," *IEEE Transactions on Neural Networks*, Vol. 3, pp. 698-713.

Carpenter, G. A., Martens, S., and Ogas, O. J., 2004, "Self-organizing Hierarchical Knowledge Discovery by an ARTMAP Image Fusion System," Proceedings, *7th International Conference on Information Fusion*, International Society of Information Fusion, Fairborn, Ohio, Vol. 1, pp. 235-242.

Fisher, R. A., 1936, "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, Vol. 7, pp. 179-188.

Granger, E., Rubin, M. A., Grossberg, S., and Lavoie, P., 2001, "A What-and-Where Fusion Neural Network for Recognition and Tracking of Multiple Radar Emitters," *Neural Networks*, Vol. 14, pp. 325-344.

Jain, A. K., Duin, R. P. W., and Mao, J., 2000, "Statistical Pattern Recognition: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, pp. 4-37.

King, R. D., Feng, C., and Sutherland, A., 1995, "StatLog: Comparison of Classification Algorithms on Large Real-World Problems," *Applied Artificial Intelligence*, Vol. 9, pp. 289-333.

Langley, P., 1994, "Selection of Relevant Features in Machine Learning," Proceedings, *AAAI Fall Symposium on Relevance*, R. Greiner and D. Subramanian, eds., AAAI Press, Los Angeles, California, pp. 140-144.

Mannan, B., Roy, J., and Ray, A. K., 1998, "Fuzzy ARTMAP Supervised Classification of Multi-Spectral Remotely-Sensed Images," *International Journal of Remote Sensing*, Vol. 19, pp. 767-774.

Michie, D., Spiegelhalter, D. J., and Taylor, C. C., 1994, *Machine Learning, Neural and Statistical Classification*, D. Michie, D.J. Spiegelhalter, and C.C. Taylor, eds., Ellis Horwood, p. 290.

Murshed N. A., Bortolozzi F., and Sabourin R., 1997, "A Cognitive Approach to Signature Verification," *International Journal of Pattern Recognition and Artificial Intelligence*, Special issue on Bank Cheques Processing, Vol. 11, pp. 801-825.

Parsons, O. and Carpenter, G. A., 2003, "ARTMAP Neural Networks for Information Fusion and Data Mining: Map Production and Target Recognition Methodologies," *Neural Networks*, Vol. 16, pp. 1075-1089.

Streilein, W., Waxman, A., Ross, W., Liu, F., Braun, M., Fay, D., Harmon, P., and Read, C. H., 2000, "Fused Multi-Sensor Image Mining for Feature Foundation Data," Proceedings, *3rd International Conference on Information Fusion*, International Society of Information Fusion, Fairborn, Ohio, Vol. 1, pp. 3-18.