Classifier Combination and Feature Selection for Land-Cover Mapping from High-Resolution Airborne Dual-Band SAR Data

Reinhold Huber
Aero-Sensing Radarsysteme GmbH, Münchner Straße 20
82234 Weßling, Germany

and

Luciano V. Dutra
Instituto Nacional de Pesquisas Espaciais
12227.000 São José dos Campos, Brazil

ABSTRACT

We study feature selection and classification for a land-cover mapping task from airborne high resolution AeS-1 data in radar X- and P-Band. The studied feature selection methods are the well-established sequential addition of features to an initially empty feature set and genetic algorithm search, both based on statistical distance measures, and exhaustive evaluation based on actual classifier performance. It was observed, that different criteria and search strategies come up with different subsets of features. We present results of combined classifications derived from classifiers trained on different subsets of features. The considered combination strategies are product, sum, maximum and majority rules. Combination turned out to bring significant improvement. The task of discriminating two forest classes, three classes of agricultural area, two classes of built-up area and a specific class devoted to radar imaging ambiguities for a test site located in Switzerland provided a satisfying result for machine classification from radar data.

Keywords: Airborne Radar, Feature Selection, Classification, Classifier Combination, Land Cover

1. INTRODUCTION

Two interesting questions in classification and pattern recognition : feature selection and classifier combination, which have been excessively studied over the last years, are addressed for a remote sensing data classification in this paper. The problem of feature selection can be stated as the problem of identifying the d most discriminative measurements out of a set of D potentially useful measurements, where $d \leq D$. In practice, feature selection turns out to be a crucial task, dependent on sample size, number of classes, class properties and measurement complexity. Classifier combination or ensemble methods have been studied in the fields of statistical and neural reasoning. In general, found either by observation or theoretical treatment, classifier combination methods tend to decrease the variance in estimation, which means an ensemble of classifiers usually performs better, with respect to overall classification accuracy and coefficients of agreement, than a single classifier.

We study feature selection and classification for a land-cover mapping task from airborne high resolution AeS-1 data. The AeS-1 sensor provides synthetic aperture radar (SAR) data in radar X- and P-Band, furthermore, as it is capable of interferometric operation, the magnitude of the complex phase correlation coefficient, coherence for short, can be regarded as an additional measure of information content. Figure 1 shows those three sources of information.

Due to the nature of SAR imaging, texture was found to be a valuable information source. Apart from the easily extractable coefficient of variation, usually calculated from local measurements and defined by the quotient of standard deviation s to mean $\bar{x}$, a number of alternative texture measurements, more prominent in the field of pattern recognition, could be extracted. Furthermore, texture can be treated using the so-called co-occurrence approach, adding a number of well-studied features to the problem. Consequently, when applying all those texture measurements to the basic sensor information, namely X/P-band backscatter and coherence, one can come up with high-dimensional measurement vector $\vec{f}$, in our case an $\vec{f}$ of dimension $D = 57$.

As already laid down by Hughes in 1968 [4], such a measurement complexity D performs suboptimal with limited training samples sizes, which is commonly the case in remote sensing, where field work is expensive and the environmental conditions are not always suitable for very effective data collection. Feature selection can be, and was, applied successfully to reduce the information to its relevant content and remove redundancy. The studied feature selection methods are the well-established sequential addition of features to an initially empty feature set and genetic algorithm search, both based on statistical distance

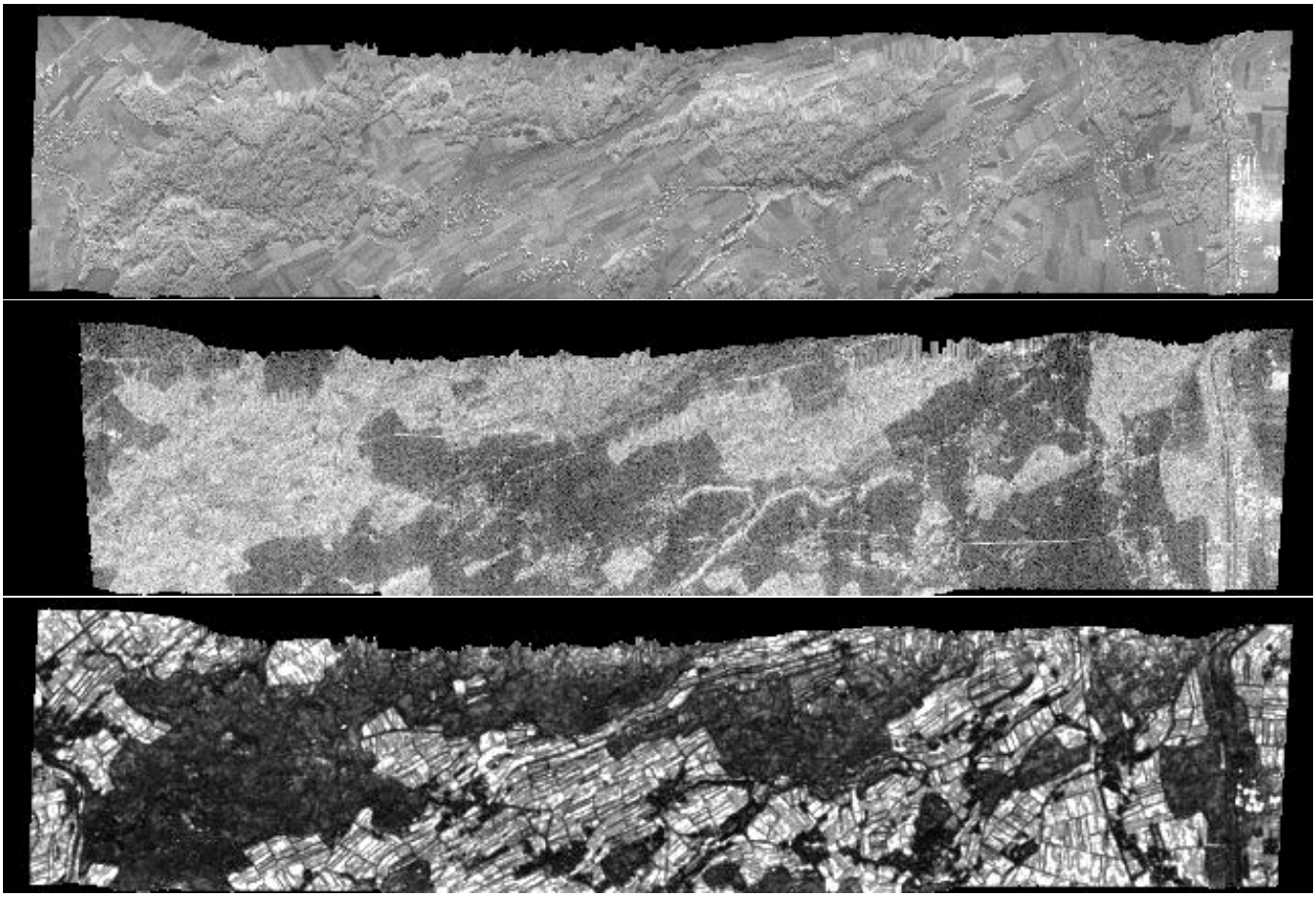

Figure 1. X–Band amplitude, P–Band amplitude and X–Band coherence.

measures, and exhaustive evaluation based on actual classifier performance. It was observed, that different criteria and search strategies come up with different subsets of features.

The latter observation and the used classification strategy, namely artificial neural networks, suggested an investigation in classifier combination (classifier is understood here as a specific combination of feature subsets and a certain classifier model). As neural networks are known to be classifiers of low bias and high variance and the found subsets are in general different with respect to the included features, classifier combination originally suggested for classifiers designed on different d-dimensional subspaces of the D–dimensional feature space, seemed to be applicable. Hence, we finally present results of combined classifications derived from classifiers trained on different portions of the feature space. The considered combination strategies are those recently suggested by Kittler [6][7], namely product, sum, maximum and majority rules. Combination turned out to bring significant improvement, as individual classifiers seem to form a so-called mixture of experts in our case.

The task of discriminating two forest classes, three classes of agricultural area, two classes of built-up area and a specific class devoted to radar imaging ambiguities for a test site located in Switzerland provided a satisfying result for machine classification from radar data.

2. FEATURE EXTRACTION

The study area is the Solothurn area in Switzerland. The data taken in X– and P–Band was geocoded to a grid width of 2.5 meters. Initial investigations of the data suggested the possible discrimination of 8 classes. Training samples were taken for these classes, namely for:

low vegetation (ω_1),
medium vegetation (ω_2),
high vegetation (ω_3),
low forest (ω_4),
high forest (ω_5),
residential area (ω_6),

industrial area (ω_7),
foreshortening (ω_8).

All classes are selected from the radar interpreters point of view, e.g. low and high forest means that there is some relation to low or high P–Band amplitude, which in turn was shown to correlate with forest biomass [11]. The class foreshortening, which actually means that information gets condensed due to specific setting of SAR imaging and surface geometry, is very apparent in forest regions. To reduce confusion with urban areas this class has to be treated separately. For specific application domains, grouping or refinement of the mentioned classes into domain specific classes might become necessary.

Statistical parameters are extracted from sliding windows over the X-,P- and coherence images. For each of which we extracted the following measures of **local statistics** [13]:

center pixel value,
mean,
coefficient of variation,
skewness,
kurtosis,
contrast,
homogeneity,
median,
range;

and from **greylevel co-occurrence matrices** [1]:

energy,
entropy,
maximum probability,
inertia,
angular second moment,
correlation,
cluster shade,
cluster prominence,
information correlation I,
information correlation II.

Application of this feature extraction process results in a feature vector $\vec{f}$ of dimension $D = 57$ for each pixel.

3. FEATURE SELECTION

Various search strategies are used to find the subset of features optimizing an adopted criterion J, once this criterion has been chosen. In the following formulation *Bhattacharryya distance* measures the separability of normal distributions for two classes indexed by i and k [2]:

$$B_{ik} = \frac{1}{8}(\vec{m}_i - \vec{m}_k)^t \Sigma^{-1} (\vec{m}_i - \vec{m}_k) + \frac{1}{2} \ln \left(\frac{|\sum|}{|\Sigma_i|^{\frac{1}{2}} \, |\Sigma_k|^{\frac{1}{2}}} \right)$$

$$\text{with} \quad \Sigma = \frac{\sum_i + \sum_k}{2}, \tag{1}$$

where $\vec{m}_i$, $\vec{m}_k$ are the feature mean vectors and Σ_i and Σ_k denote class covariance matrices for classes i and k, respectively. Both, class mean vectors and covariances have been estimated from available training data.

For multiclass problems it is appropriate to use the average *Jeffreys–Matusita Distance* (JMD) as separability criteria [5]. For C classes and equal a–priori probabilites $p(\omega_i), i = 1 \ldots C$, the average JMD is defined by:

$$J = \frac{2}{C(C-1)} \sum_{i=1}^{C} \sum_{j=1}^{i-1} J_{ik}, \qquad J_{ik} = 2(1 - e^{-B_{ik}}). \tag{2}$$

We compared two feature selection methods, the most established one, and one based on evolutionary algorithms (EAs) in our work.

3.1. Sequential Forward Selection (SFS) [14]

Initialize the algorithm by setting $X_0 \equiv \emptyset$. Suppose k features have already been selected from the complete set of measurements $Y = \{y_j | j = 1, 2, \cdots, D\}$ to form feature set X_k. The $(k+1)^{st}$ feature is then chosen from the set of available measurements, $Y \setminus X_k$, so that

$$J(X_{k+1}) = \max_{\forall y_j} J(X_k \cup \{y_j\}), \quad y_j \in Y \setminus X_k. \tag{3}$$

3.2. Genetic Algorithm based Selection (GAS) [12]

In the basic approach to feature selection using EAs a feature subset is encoded as a binary vector (*Bitmap Encoding*) $\vec{a} = (a_1, \ldots, a_d)$, where $a_i = 1$ indicates the presence of the $i-$th feature in the subset, while the absence of the $i-$th feature is expressed by $a_i = 0$. To restrict the subset to a specific size *permutation encoding* is better suited than bitmap encoding. In permutation encoding each chromosome is represented as a permutation as shown in Figure 2. In order to preserve

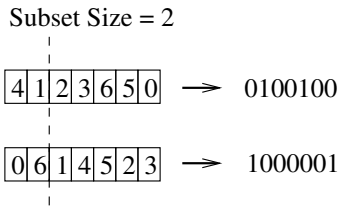

Figure 2. Permutation encoding.

the permutation property of the chromosomes, specific genetic operators have been devised. The mutation operator simply exchanges two random bases on the chromosome with a given mutation rate p_m (usually in the range of $0.001 - 0.01$). One of the most prominent crossover operators for permutation encoding is the *Partially Matched Crossover* (PMX) proposed in [3]. Its basic mechanisms are presented in Figure 3. Generally, for crossover two parent chromosomes are selected, and crossover

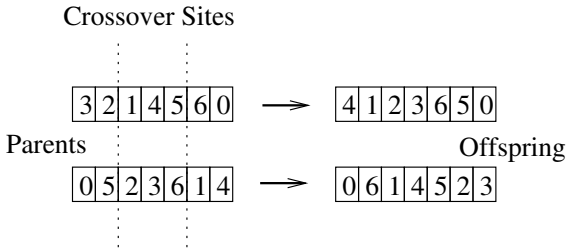

Figure 3. Partially Matched Crossover (PMX).

is performed according to a user–defined crossover rate p_c (usually in the range of $0.6 - 1.0$). If no crossover occurs, the two parents are simply copied to two offspring chromosomes. In the crossover phase two crossover sites are selected randomly (sites are the same for both parents). Then, the bases in between the crossover sites are exchanged. Up to this point we have exactly described the very common *2–point Crossover*, but if the bases were only exchanged, the permutation property would be lost. Thus, each base to be copied to the other parent is searched in that parent and swapped with the base currently at the locus, where the exchange takes place (just like a single mutation). In doing so the partial order between the crossover sites can be exchanged between the parents without corrupting the permutation property.

4. CLASSIFIER COMBINATION

In the following, $p(\omega_j)$ denotes the a-priori probability of class ω_j and $p(\vec{x}_i|\omega_j)$ is the class conditional probability density function for a realization $\vec{x}_i$ of $\vec{f}$.

Assuming statistical independence for the feature vectors $\vec{x}_i$ of classifier i, the **Product Decision Rule** for class ω_j becomes:

$$p(\omega_j)\Pi_{i=1}^{R} p(\vec{x}_i|\omega_j) = \max_{k=1}^{C} p(\omega_k)\Pi_{i=1}^{R} p(\vec{x}_i|\omega_k), \tag{4}$$

where R classifiers are fused and the decision is made among C classes. In terms of a–posteriori probabilities the product rule can be written as [8]:

$$p^{-(R-1)}(\omega_j)\Pi_{i=1}^{R} p(\omega_j|\vec{x}_i)p(\vec{x}_i) = \max_{k=1}^{C} p^{-(R-1)}(\omega_k)\Pi_{i=1}^{R} p(\omega_k|\vec{x}_i)p(\vec{x}_i). \tag{5}$$

Under the rather strong assumption that posterior and prior probabilities are similar, i.e. $p(\omega_k|\vec{x}_i) = p(\omega_k)(1 + \delta_{ki})$ with $\delta_{ki} << 1$, the **Sum Decision Rule** for class ω_j (after product expansion and elimination of high order terms) becomes:

$$(1 - R)p(\omega_j) + \sum_{i=1}^{R} p(\omega_j|\vec{x}_i) = \max_{k=1}^{C}\Big((1 - R)p(\omega_k) + \sum_{i=1}^{R} p(\omega_k|\vec{x}_i)\Big). \tag{6}$$

Under the assumption of equal a–priori class probability , the **Maximum Decision Rule** is derived. Replacing the sum in Equation (6) by a upper bound given by $\frac{1}{R}\sum_{i=1}^{R} p(\omega_k|\vec{x}_i) \leq \max_{i=1}^{R} p(\omega_k|\vec{x}_i)$ we assign class ω_j, if:

$$\max_{i=1}^{R} p(\omega_j|\vec{x}_i) = \max_{k=1}^{C} \max_{i=1}^{R} p(\omega_k|\vec{x}_i). \tag{7}$$

The **Majority Vote Rule** operates on a binary valued function Δ_{ki}, where $\Delta_{ki} = 1$ if $p(\omega_k|\vec{x}_i) = \max_{j=1}^{C} p(\omega_j|\vec{x}_i)$ and $\Delta_{ki} = 0$ otherwise, we assign class ω_j by majority voting through:

$$\sum_{i=1}^{R} \Delta_{ji} = \max_{k=1}^{C} \sum_{i=1}^{R} \Delta_{ki}. \tag{8}$$

5. RESULTS

Five strategies for feature selection have been applied. A subset size of 5, which is the number of basic features devised in SAR image analysis literature, was used to compare those suggested features to features found by SFS and GAS. These three experiments were based on evaluation of the average JMD value. With feature subsets of moderate size an exhaustive evaluation of subsets becomes possible. The number of necessary evaluations is given by 2^D. Clearly, for $D = 57$ this is impossible. Furthermore, it is favorable to use the classifier accuracy as a performance measure. Therefore, we used SFS and GAS based on JMD to produce subsets of dimension 8, which we further refined by exhaustive search on classifier accuracy.

Standard features are X–Band amplitude, X–Band coefficient of variation, P–Band amplitude, P–Band coefficient of variation, X–Band coherence.

SFS for 5 best features yielded X–Band mean, skewness and inertia, P–Band mean and contrast.

GAS for 5 best features resulted in X–Band mean, contrast and cluster prominence for coherence, P–Band cluster prominence.

SFS/exhaustive search SFS for 8 best features followed by exhaustive classifier accuracy evaluation produced the 7–dimensional subset consisting of X–Band mean, coefficient of variation, skewness, range, energy and inertia, P–Band information correlation II.

GAS/exhaustive search GAS for 8 best features followed by exhaustive classifier accuracy evaluation selected an 5–dimensional subset including X–Band mean, contrast and inertia, P–Band mean, contrast of coherence.

The used Artificial Neural Network (ANN) classifiers were single–hidden layer perceptrons trained with resilient backpropagation [10]. The networks were fully connected between input and hidden layer, and hidden and output layer, respectively. The hidden layer was fixed to 10 neurons, which was heuristically found to be a lower bound, below which accuracy tended to decrease caused by too low network capacity. Training was stopped as mean squared error started to stabilize [9].

We present ANN classification confusion matrices, where the rows have to be read as the reference categories. The columns correspond to the ANN classifier decision, i.e. the class assignment based on a winner–takes–all decision.

Standard features accuracy = 0.8038

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
ω_1	192	24	1	11	1	3	2	0
ω_2	16	231	8	4	0	0	0	0
ω_3	0	10	249	7	0	0	0	0
ω_4	6	4	9	288	7	50	0	12
ω_5	0	0	0	16	218	5	3	15
ω_6	2	0	0	42	6	191	5	3
ω_7	0	0	0	9	6	16	155	8
ω_8	0	0	14	34	34	4	3	119

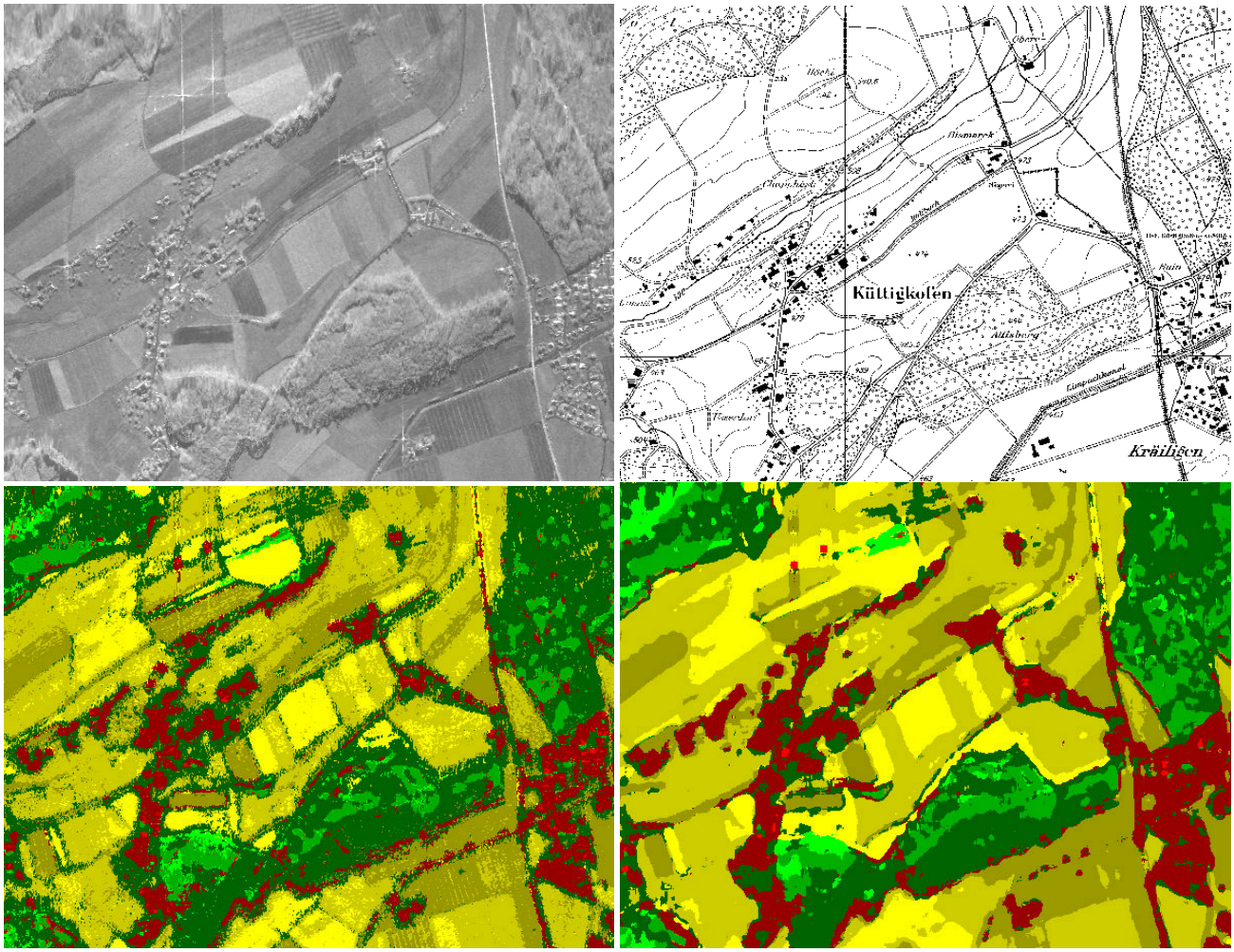

Figure 4. Detailed View of X–band, Scanned Map, Classification Based on Standard Features and Combined Classification.

SFS / exhaustive search accuracy = 0.9090

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
ω_1	216	8	1	3	2	4	0	0
ω_2	2	249	8	0	0	0	0	0
ω_3	0	1	256	2	0	0	0	8
ω_4	0	0	0	352	7	9	0	8
ω_5	0	0	0	7	235	0	2	13
ω_6	0	0	0	28	5	214	1	1
ω_7	0	0	0	15	3	4	167	5
ω_8	0	0	1	14	22	1	2	169

We will not list all results for classifiers based on different selection schemes in detail. For short, the achieved accuracy for all other selection methods is on the order of magnitude of the SFS/exhaustive search approach. As the standard features approach is approximately 10 percents below of any single best classifier, combination is based on ANNs trained on subsets originating from the remaining 4 selection schemes.

Sum Decison Rule accuracy = 0.9310

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
ω_1	218	7	1	1	1	6	0	0
ω_2	0	251	8	0	0	0	0	0
ω_3	0	1	256	0	0	0	0	8
ω_4	0	0	0	359	6	4	0	8
ω_5	0	0	0	4	242	1	2	13
ω_6	2	0	0	7	1	237	0	1
ω_7	0	0	0	11	6	6	167	5
ω_8	0	0	1	13	23	0	0	171

Product Decison Rule accuracy = 0.9110

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
ω_1	217	7	1	1	0	8	0	0
ω_2	0	252	7	0	0	0	0	0
ω_3	0	2	257	0	0	0	0	8
ω_4	0	0	0	358	7	4	0	7
ω_5	0	0	0	5	237	1	2	12
ω_6	2	0	0	20	4	223	0	2
ω_7	0	0	0	11	7	7	164	5
ω_8	10	0	3	13	23	0	5	154

Maximum Decison Rule accuracy = 0.9291

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
ω_1	221	6	1	0	1	5	0	0
ω_2	0	250	8	0	0	1	0	0
ω_3	0	1	258	1	0	0	0	7
ω_4	3	0	0	360	6	2	0	5
ω_5	0	0	0	8	237	1	2	9
ω_6	5	0	0	12	1	231	0	0
ω_7	1	0	0	7	5	6	172	3
ω_8	0	0	1	14	20	0	3	170

Majority Vote Rule accuracy = 0.9237

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
ω_1	219	5	1	2	1	6	0	0
ω_2	0	252	7	0	0	0	0	0
ω_3	0	1	258	0	0	0	0	8
ω_4	0	0	0	363	6	2	0	5
ω_5	0	0	0	7	242	0	0	8
ω_6	3	0	0	18	2	224	0	2
ω_7	0	0	0	13	6	5	166	4
ω_8	0	0	1	18	24	0	1	164

Figure 5 shows results for standard feature based classification, for classification based on SFS followed by exhaustive search and combined classifiction using the sum decison rule.

6. CONCLUSION

Two important observations made in recent literature have been experimentally verified for radar data: textural features subsets of moderate size perform well or even better than large sets (given a limited sample of training data), combination of classifiers often outperform single classifiers. The considered number of 8 classes is rather large and helps in automatic map production, though manual inspection is still required for high–quality map production. Further directions of research are the study of other feature selection algorithms and separabibility criterias. Classifier combination based on error correcting combination and different classifier models (e.g. Gaussian classifiers operating on class mixture models instead of ANNs), where classifiers are weighted by their performance on specific classes is another ingredient towards further improvement of classifcation accuracy.

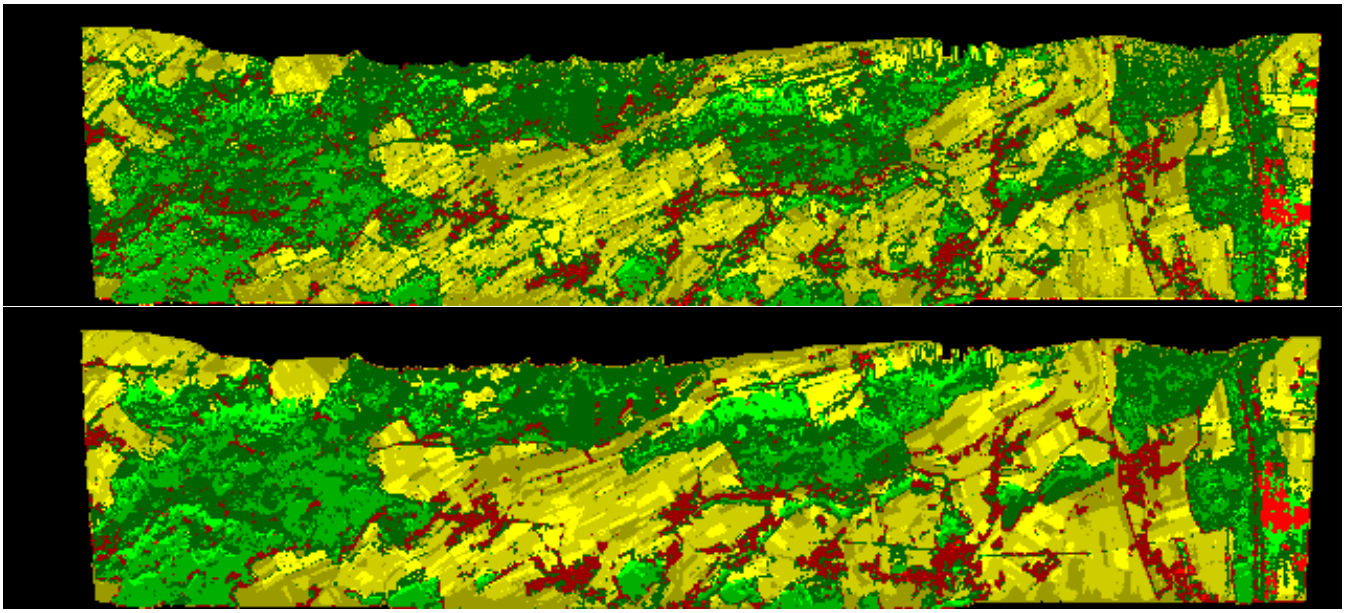

Figure 5. Classification Using Standard Features and Combined Classifcation Using Sum Rule.

7. REFERENCES

[1] D. DeKruger and B.R. Hunt. Image Processing and Neural Networks for Recognition of Cartographic Area Features. *Pattern Recognition*, 27(4):461–483, 1994.

[2] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[3] David E. Goldberg. *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison–Wesley, 1989.

[4] Gordon F. Hughes. On the Mean Accuracy of Statistical Pattern Recognition. *IEEE Transactions on Information Theory*, 14(1):55–63, January 1968.

[5] Thomas Kailath. The Divergence and Battacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communications*, 15(1):52–60, 1967.

[6] Josef Kittler. Combining Classifiers: A Theoretical Framework. *Pattern Analysis & Applications*, 1(1):18–27, 1998.

[7] Josef Kittler, Mohamad Hatef, Robert P.W. Duin, and Jiri Matas. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

[8] Tong Lee, John A. Richards, and Philip H. Swain. Probabilistic and Evidential Approaches for Multisource Data Analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 25(3):283–293, May 1987.

[9] Lutz Prechelt. Automatic Early Stopping using Cross Validation: Quantifying the Criteria. *Neural Networks*, 9(3):457–462, April 1996.

[10] Martin Riedmiller and Heinrich Braun. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In *Proceedings of International Conference on Neural Networks*, pages 586–591, San Francisco, CA, USA, 1993.

[11] Eric J. Rignot, Reiner Zimmermann, and Jakob J. van Zyl. Spaceborne Applications of P Band Imaging Radars for Measuring Forest Biomass. *IEEE Transactions on Geoscience and Remote Sensing*, 33(5):1162–1169, September 1995.

[12] Wojciech Siedlecki and Jack Sklansky. A Note on Genetic Algorithms for Large–Scale Feature Selection. *Pattern Recognition Letters*, 10:335–347, 1989.

[13] Anne H. Schistad Solberg and Anil K. Jain. Texture Fusion and Feature Selection Applied to SAR Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 35(2):475–479, March 1997.

[14] A. Wayne Whitney. A Direct Method of Nonparametric Measurement Selection. *IEEE Transactions on Computers*, 20:1100–1103, September 1971.