# Application of Bayesian Statistical Anaylsis to Multisource Data Integration

Sa-hyun, Hong[1], Wooil M. Moon[1,2]
1, ESI3 Laboratory, School of Earth and Environmental Sciences, Seoul National University, Kwan-ak Gu Shil-rim dong san 56-1, Seoul, 151-742, Korea
(shhong@eos1.snu.ac.kr, wmoon@eos1.snu.ac.kr)
Tel : +82-2-880-6526, Fax : +82-2-871-3269
2, Geophysics, The University of Manitoba, Winnipeg, Canada R3T 2N2
(wmoon@cc.umanitoba.ca)

## Abstract

In this paper, Multisource data classification methods based on Bayesian formula are considered. For this decision fusion scheme, the individual data sources are handled separately by statistical classification algorithms and then Bayesian fusion method is applied to integrate from the available data sources. This method includes the combination of each expert decisions where the weights of the individual experts represent the reliability of the sources. The reliability measure used in the statistical approach is common to all pixels in previous work. In this experiment, the weight factors have been assigned to have different value for all pixels in order to improve the integrated classification accuracies. Although most implementations of Bayesian classification approaches assume fixed *a priori* probabilities, we have used adaptive *a priori* probabilities by iteratively calculating the local *a priori* probabilities so as to maximize the *posteriori* probabilities. The effectiveness of the proposed method is at first demonstrated on simulations with artificial and evaluated in terms of real-world data sets. As a result, we have shown that Bayesian statistical fusion scheme performs well on multispectral data classification.

## I. INTRODUCTION

Data fusion means an approach to the integration of information from several different sources, aiming at an improved quality of results with mathematical framework[1]. We are interested in using all the available data sources to extract more information and obtain higher classification accuracy. There are many developed data fusion methods: Dempster-Shfer evidential reasoning[2], Neural networks[3] and Fuzzy concepts[4] though, our attention is focused on Bayesian statistical method proposed by Benediktsson *et al.*[5]. In this statistical case, the relative reliabilities of the sources involved in classification must be taken into account. This requires a way to characterize and quantify the reliability of a data source, which becomes important when we apply the scheme to multisource data classification. A number of previous studies [5][6][7] have been focused on estimation of overall reliability for each source, not being taken the deviations of data source sensitivity to classes into consideration. We will investigate methods to determine the reliability factors and to translate them into weights to be used in the classification process.

## II. STATISTICAL DATA INTEGRATION

In the general multisource data fusion case, we have a set of images from $n$ separate data sources, each providing a feature vector $x_s$, $s = 1,\ldots, n$ for each of the pixels of interest. Assuming that the study area consists of $M$ information classes denoted by $w_j$, $j = 1,\ldots, M$ into which the image's pixels are to be classified and let the $i_{th}$ data class from the $S_{th}$ source be denoted by $d_{si}$, $i = 1,\ldots, m_s$, where $m_s$ is the total number of data classes for the source.

Now suppose that the estimation about information classes can be considered from the

collections of data-specific classes, which means mathematically that the information classes of interest $w_j$ are related to the data classes for a single source by means of a set of source-specific decision functions $f(w_j \mid d_{si}(x_s))$ for all $i$, $j$, $s$. This function is a measure of strength of association of data class $d_{si}$ with the information class $w_j$[8].

Hence, we can formulate the global decision function related with information class $w_j$ as:

$$\Omega_j = F\{f(w_j \mid d_{si}(x_s)),\ \lambda_i \mid i = 1,..,\ m_s,\ s = 1,..,\ n\} \quad (1)$$

$$F\ :\ f(w_j,\ d_{sb},\ \lambda_i) \ \rightarrow\ \Omega_j \in [0,\ 1] \quad (2)$$

Where $\lambda_i$ is the reliability factor we pose it to the $S_{th}$ source. Provided that a pixel measurement vector $\mathbf{X} = [x_1,\ldots,\ x_n]^t$, it will be classified depending on the decision rule: Decide $\mathbf{X}$ is in class $\hat{w}$ for which[8]

$$\hat{\Omega} = \max_j \Omega_j \quad (3)$$

To implement this Bayesian statistical formula the decision function should be rewrite by Bayesian chain rule.

$$\Omega_j(\mathbf{X}) = Prob(w_j \mid \mathbf{X}) = Prob(w_j \mid x_1,\ldots,\ x_n) \quad (4)$$

$Prob(w_j \mid x_1,\ldots,x_n)$ is the posteriori probability that $w_j$ is the correct class given that the observation, $[x_1,\ldots,x_n]$ from the sources made up at $w_j$. Assuming class conditional independence among the $n$ sources, this can be set as:

$$Prob(x_1,\ldots,\ x_n \mid w_j) = Prob(x_1 \mid w_j)\cdots Prob(x_n \mid w_j) \quad (5)$$

This assumption does not always hold. Especially if the images were taken with similar wavelengths then there might well be a significant correlation between deviations in the sensor responses. However, the conditional independence assumption greatly reduces the mathematical effort required. Therefore the global decision function $\Omega_j(\mathbf{X})$ is simplified as:

$$\Omega_j(\mathbf{X}) = Prob(w_j \mid x_1)\cdots Prob(w_j \mid x_n)\ Prob(w_j)^{1-n} \quad (6)$$

We can express the individual source specific decision function as:

$$Prob(w_j \mid x_s) = \sum_{i=1,m_s} \frac{Prob(w_j, d_{si}, x_s)}{Prob(x_s)} \quad (7)$$

which leads to

$$Prob(w_j \mid x_s) = \sum_{i=1,m_s} Prob(w_j \mid d_{si}, x_s) Prob(d_{si} \mid x_s) \quad (8)$$

If we consider the source reliability factor $\lambda$, the global decision function is

$$\Omega_j(\mathbf{X}) = Prob(w_j) \prod_{s=1}^{n} \left\{ \frac{Prob(w_j \mid x_s)}{Prob(w_j)} \right\}^{\lambda_s} \quad (9)$$

where $\lambda_s\ (s = 1,\ldots,\ n)$ ranges in [0,1]. Note that $\lambda_s = 0$ means $S_{th}$ source is totally unreliable and $\lambda_s = 1$ means $S_{th}$ source is the most reliable

However, this method of putting exponents on the probabilities does not influence on the decision process because the exponential function $Prob^{\lambda}$ is a monotonic function of $Prob$. In order to integrate multisource decisions here, we consider the logarithmic form as[5]

$$log\ \Omega_j(\mathbf{X}) = log\{Prob(w_j)\} + \sum_{s=1}^{n} \lambda_s\ log \left\{ \frac{Prob(w_j \mid x_s)}{Prob(w_j)} \right\} \quad (10)$$

In the current research, the source reliability is evaluated by a function of the known probability of detection $P_D$ and the probability of false alarm $P_F$

$$W = log(Prob_D / Prob_F) \quad (11)$$

All the pixels in the scene therefore have different calculated reliabilities for each source $S_n$.

## III. TEST EXAMPLES

To implement the global discrimination function Eq(10) we need a simple algebraic manipulation, and Eq(10 ) can be recast in the form

$$Prob(w_j \mid x_s) = \sum_{i=1,\ m_s} \frac{Prob(x_s \mid d_{si}, w_j) Prob(d_{si}, w_j)}{Prob(x_s)} \quad (12)$$

$$Prob(x_s) = \sum_{j=1,M} \sum_{i=1,m_s} Prob(x_s \mid d_{si}, w_j) Prob(d_{si}, w_j) \quad (13)$$

where $M$ is the number of information classes. By assuming the conditional independence, we have

$$Prob(x_s \mid d_{si}, w_j) = Prob(x_s \mid d_{si}) \quad (14)$$

*Simulated Data Set*

First, synthetic images with different contrast are used to illustrate the applicability of our method. Two images corrupted by gaussian noise are shown in Figure 1.

Each image contains four regions(information classes, M = 4). These images are combined in order to provide a classification of the image into four classes. The integrated output images are shown in Figure 2 where M classes are well identified and discriminated. Figure 2-(b) represents that the utilizing information about the local homogeneity could significantly increase the a posteriori classification accuracy by changing the a priori probabilities [9]

*Real Data Set*

Application of the Bayesian probabilistic method with considered weighting scheme for the integration of multisource image data is examined with a real high dimensional data set by decomposing the data into small pieces, i.e., subsets of spectral bands. The data set used in this experiment is MASTER data obtained by PACRIM II Campaign, 2000. Table 1 provides a description of the MASTER data set.

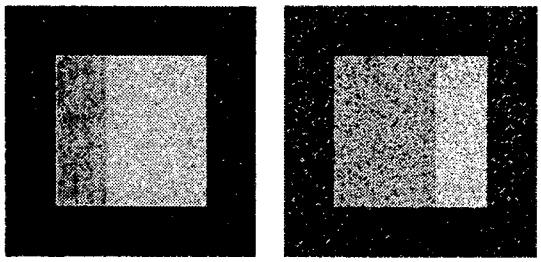
Figure 3. is a visual representation of global statistical correlation coefficient matrix of the data. Based on the correlation image, the 50 channels were divided into two groups in such a way that intra-correlation is maximized and inter-correlation is minimized. Table 2. describes the multisource data set after division.

The study area has 6 information classes that listed in table. 3. In our research, the process of determining the reliability factors was characterized as several schemes such as an Identical Weighting(IW), Classification Accuracy of a data source(CA) and a ratio of probability of Detection to probability of False alarm for each pixels(DFR). Figure 4 shows the Maximum likelihood(ML) classification results of data source 1 and source 2, respectively for decision level data fusion.

The right portion of the classification image of source 2 is misclassified dominantly into housing category being represented by blue color. That is because the spectral features[0.46 $\mu m$ – 2.40 $\mu m$] consisting of source 2 is partially marred by clouds. Figure 5. shows the classification results of statistical integrated images according to various determining process of weight factors. By experimental consequences, our proposed pixel-by-pixel weighting scheme shows the best performance over the multispectral data integration. The quantitative classification agreement with training samples and test samples is reported in Table 4.

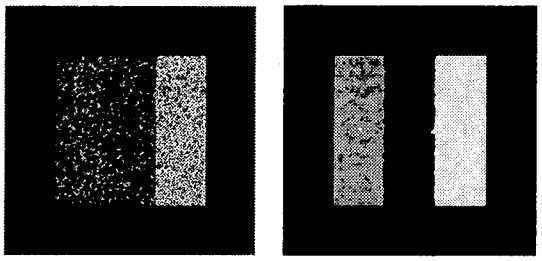
## IV. DISCUSSION AND CONCLUSION

In this paper we have investigated how Bayesian statistical probabilistic method can be used to aggregate information from various independent data sources. The logarithmic form of statistical discrimination function performed well in the classification of high dimensional data. In the probabilistic case the experiments demonstrate that the reliability measure used in data fusion have significant influence on the final fused results. The classification using two data sources associated with our suggested method regarding the selection of weights pixel-by-pixel basis gave improvement in overall classification accuracies as compared to other weighting methods(*Figure 6*).



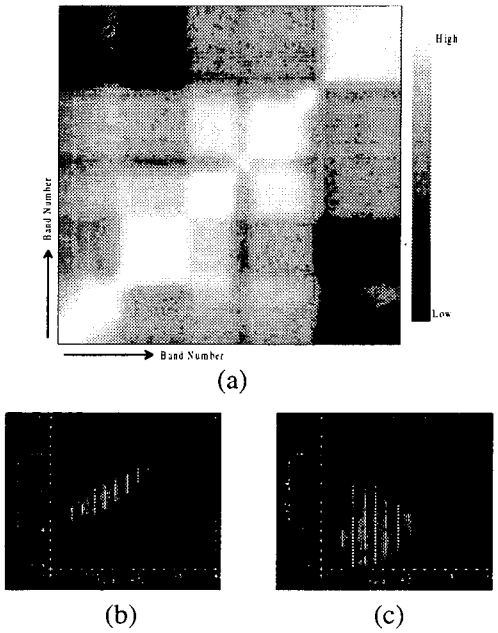
(a) (b)

*Figure 1: Two images simulating with different statistical parameters. (a) Image of the first source (b) Image of the second source*



(a) (b)

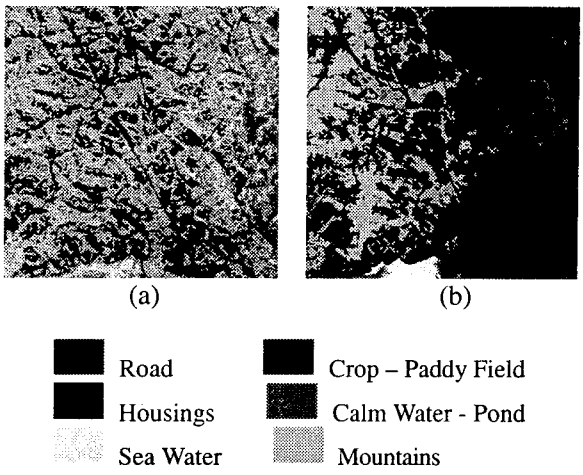
*Figure 2: Ouput integrated images obtained with the proposed Bayesian statistical fusion scheme. (a) Fused result with assuming fixed a priori probabilities (b) Fused result with pixel spatial information by local neighborhood probabilities*




(a)

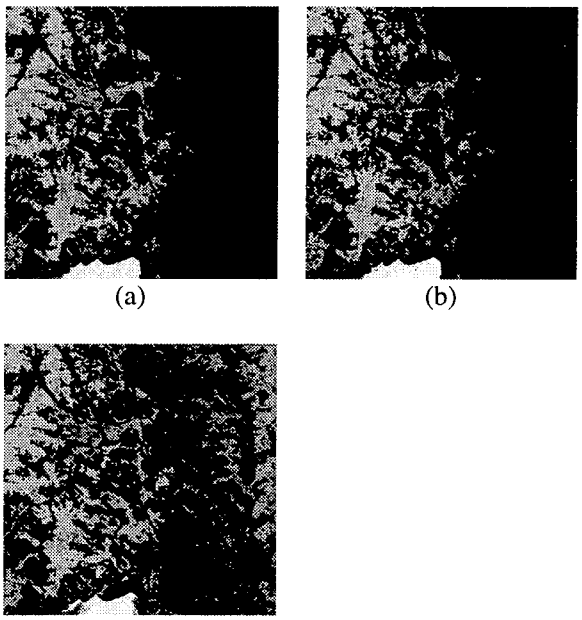
(b) (c)

*Figure 3.(a) Global statistical correlation coefficient image of the data set. The lighter the brightness, the more correlated are the spectral bands (b), (c) are intra-source scatterplot(highly correlated) and inter-source scatterplot (uncorre - -lated)*




(a) (b)

*Figure 4. Classification results of individual data source (a) Classification image of Source 1 (b) Classification image of Source 2. The right part of the image (b) is classified into Housing class due to spectral modulation incurred by cloud cover.*



(a) (b)

(c)

*Figure 5. Classification images for several determining process of weighting factors (a) Identical Weighting for each source (b) Weighted by Classification Accuracy for each source (c) Our suggested weighting scheme by a ratio of detection to false alarm.*

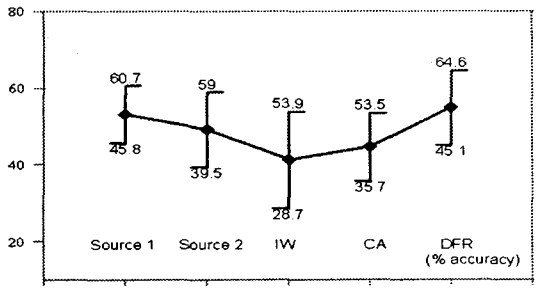



*Figure 6. Classification accuracy variations depending on the weighting scheme*

*Table .1 Description of MASTER Data Set.*

| Name | GongJu MASTER Data Set |
|---|---|
| Data Type | 50-bands Hyperspectral Data |
| Spectral Region | 0.4620 – 12.1650 ( $\mu m$ ) |
| Image size | 500 samples ×500 lines |

*Table 2. Divided Sources of High Dimensional MASTER Data Set.*

| Source Index | Input Channels | No. of Features |
|---|---|---|
| Source 1 | 23[8.225 $\mu m$ ] - 50[12.17 $\mu m$ ] | 28 |
| Source 2 | 1[0.462 $\mu m$ ] – 22[2.394 $\mu m$ ] | 22 |

*Table 3. Information Classes in the Experiment on the Multisource Data Set*

| Class No. | Information Class | Training Samples |
|---|---|---|
| 1 | Road/Highway | 191 |
| 2 | Crop/paddy field | 436 |
| 3 | Housing/Small village | 472 |
| 4 | Coastal Sea Water | 505 |
| 5 | Mountains | 575 |
| 6 | Calm Water/Lake | 424 |

*Table 4. Classification Results For Training Samples and Test Samples(S1: Source 1, S2: Source 2, IW: Identical Weighting, CA: Weighted by Classification Accuracy, DFR: Weighted by Detection to False alarm Ratio). Testing pixels are sampled mostly in the interrupted area by clouds*

| | Percent Classification Accuracy for Classes | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | With Training Samples(3073 pixels) | | | | | With Test Samples(1733 pixels) | | | | |
| | S1 | S2 | IW | CA | DFR | S1 | S2 | IW | CA | DFR |
| 1 | 88.7 | 100 | 82.3 | 96.5 | 99.1 | 93.9 | 100 | 20.2 | 86.3 | 100 |
| 2 | 64.1 | 98.6 | 98.6 | 86.7 | 98 | 16.4 | 10.1 | 10 | 0 | 34.8 |
| 3 | 86.6 | 98.3 | 98.3 | 93.4 | 98 | 12.7 | 58.5 | 100 | 99.7 | 37.8 |
| 4 | 63.2 | 100 | 100 | 99 | 100 | 73.8 | 99.3 | 94.9 | 94.9 | 57.6 |
| 5 | 39.3 | 100 | 100 | 98.1 | 100 | 84.2 | 34.4 | 0 | 0 | 0 |
| 6 | 49.9 | 99.7 | 91.4 | 91 | 100 | 10.8 | 12.5 | 0 | 0 | 99.3 |
| OA | 62.4 | 99.4 | 97 | 94.3 | 99.3 | 45.8 | 39.5 | 28.7 | 35.7 | 45.1 |

ACKNOWLEDGEMENT

This study is partially funded by the BK21 program through School of Earth and Environmental Sciences(SEES), Seoul National University and partially by NSERC operating grant(A-7400) to Wooil M. Moon.

REFERENCES

[1] L. Wald, "A Conceptual Approach to the Fusion of Earth Observation Data", *Surveys in Geophysics* 21: 177-186, 2000

[2] Sylvie Le Hégarat-Mascle, Isabelle Bloch and D. Vidal-Madjar, "Application of Dempster-Shafer Evidence Theory to Unsupervised Classification in Multisource Remote Sensing", *IEEE-Trans. Geosci. Remote Sensing* Vol.35, No.4, July 1997

[3] Jon. A. Benediktsson and Ioannis Kanellopoulos, "Classification of Multisource and Hyperspectral Data Based on Decision Fusion", *IEEE-Trans. Geosci. Remote Sensing,* Vol.37, N0.3, May 1999

[4] P. K. Hou, X. J. Wang, X. Z. Shi, L. J. Lin and M. Z. Zhang, "Target Recognition Using Fuzzy Fusion Classifier", *International Symposium on Information Fusion(ISIF)*, 2000

[5] Jon A. Benediktsson and Philip H. Swain, "Consensus Theoretic Classification Methods", *IEEE-Trans. Geosci. Remote Sensing,* Vol.22, No.4, July/August 1992

[6] Jon A. Benediktsson, Philip H. Swain and Okan K. Ersoy, "Neural Network Approaches Versus Statistical Methods in Classification of Multisource Remote Sensing Data", *IEEE-Trans. Geosci. Remote Sensing*, Vol.28, No.4, July 1990

[7] H. Kim and P. H. Swain, "A method for classification of multisource data using interval-valued probabilitied and its application to HIRIS data", *Multisource Data Integration in Remote Sensing*, NASA Conf. Publication 3099, 1991

[8] T. Lee, J. A. Richards and P. H. Swain, "Probabilistic and Evidential Approaches for Multisource Data Analysis", *IEEE-Trans. Geosci. Remote Sensing*, Vol.25, 1987

[9] J. J. Van Zyl and C. F. Burnette, "Bayesian classification of polarimetric SAR images using adaptive a priori probabilities", *Int. J. Remote Sensing*, Vol.13, No.5, 1992