부분 훈련 인식을 위한 최적 중요도 검증기법

정회원 전 병 우*, David A. Landgrebe**

Optimal Significance Testing for Partially Supervised Classification

Byeungwoo Jeon*, David A. Landgrebe** *Regular Members*

ABSTRACT

Significance testing is one of the most widely used techniques in various applications of statistical analysis. We apply significance testing to the partially supervised classification problem in which one is interested in identifying only a particular class of interest, based on the class definition through training samples of that class. One important element in applying significance testing to classification is the significance level, which should be provided by the data analyst in such a way that omission or type I error is limited at a pre-specified level. This paper addresses the problem of unsupervised estimation of an optimal significance level using the class-averaged and generalized total classification error criteria, and applies its result to partially supervised classification.

I. 서 론

Successful classification of a given data set requires proper design of classifiers to be employed. To be maximally effective, the design of a classifier requires prior information which is usually given in the form of training samples. The number of training samples is dependent on the number of features and the number of classes[1,2]. In practical applications of pattern classification techniques, a frequently observed characteristic is the heavy, often nearly impossible requirement on representative prior statistical characteristics of all classes in a given data set.

This paper deals with designing a classifier that can identify a particular class of samples with statistical information pertaining only to that class of interest. This kind of problem is especially important where defining all the classes and gathering corresponding statistical information is impossible or very expensive in terms of time and manpower. We call this a "partially supervised classifier"[3] in the sense that the prior information is available only for the class of interest, thus partially supervised. Classifiers such as the parallelepiped classifier[4] or a scheme based upon a known absorption feature for a specific material identify samples on an absolute basis, that is, without relative comparison to other alternatives. In such cases, class definition through training samples is required only for the particular class of interest. This kind of problem is also known as the single hypothesis problem[2] or one-class classifier[5].

Significance testing is a widely used technique in various applications of statistical analysis[6]. A partial list of examples includes target detection, object detection out of various backgrounds[7], texture detection, cloud detection, and fault or anomaly detection in diagnostic monitoring[8]. In this paper, we apply significance testing to the partially supervised classification problem.

One of the important elements in significance testing is the significance level which must be provided by the data analyst usually in such a way that the omission

* School of Electrical and Computer Engineering, Sung Kyun Kwan University, Suwon, Korea, (bjeon@yurim.skku.ac.kr)
** School of Electrical Engineering, Purdue University, W. Lafayette, IN 47907-1285, U.S.A., (landgreb@ecn.purdue.edu)

논문번호 : 98557-1231, 접수일자 : 1998년 12월 31일

(*i.e.*, type I) error is kept within a pre-specified level[9]. From an application point of view, it will be useful if one can use other criteria, such as the Bayes minimum error criterion, or the ones used in the minimax, or Neyman-Pearson testing[10] to determine a suitable significance level. Unfortunately, lack of prior statistical information other than that of the particular class of interest prevents evaluating the commission (or, type II) error, and thus, forbids directly applying conventional procedures used in hypothesis testing.

Motivated by the notion that one can predict the commission error using the mixture density estimate of the selected test statistic, this paper presents a method which estimates the optimal significance level using the mixture density of the selected test statistic estimated from the given (unlabeled) data set. Note that the test statistic is one-dimensional and what needs to be estimated is the *mixture* density. As optimality conditions, one can use Bayes total classification error, minimum class-averaged classification error, or the generalized total classification error criteria. Its result is used in partially supervised classification.

II. Significance Testing Applied To Partially Supervised Classification

Suppose there is a data set, $\mathbf{X} \equiv \{x_1, \cdots, x_N\}$ with N samples. Each data sample, x_i, is a q-dimensional feature vector ($q \geq 1$). We assume that one is interested in identifying only a single class (denoted by C_{int}), that is, discriminating between it and the others class (denoted by C_{others}). The others class might consist of several subclasses none of which are of one's interest. Let $f_x(x|C_{int})$ and $f_x(x|C_{others})$ be the probability density functions of classes C_{int} and C_{others}, and let corresponding prior probabilities be indicated respectively by π_{int} and π_{others}. Prior statistical knowledge is assumed to be available only for the class of interest, thus only $f_x(x|C_{int})$ is known. The mixture probability density, denoted by $f_x(x)$, is written as, $f_x(x)= \pi_{int} f_x(x|C_{int}) + \pi_{others} f_x(x|C_{others})$ where $0 < \pi_{int}, \pi_{others} < 1$, $\pi_{int} + \pi_{others} = 1$. Even though the derivations henceforth do not require any specific family of probability density functions for C_{int}, for simplicity's sake, multivariate normality is assumed for C_{int}. Generalization to other probability density functions is straightforward. Furthermore, without loss of generality, C_{int} is assumed to have zero mean, denoted by $\mathbf{O}_q$, and an identity covariance matrix, denoted by $\mathbf{I}_{qxq}$. This standard multivariate normal distribution is denoted by $MVN[\mathbf{O}_q, \mathbf{I}_{qxq}]$.

In the partially supervised classification using significance testing, a single hypothesis $H_1 : x \in C_{int}$, is tested against all other alternatives to identify samples belonging to the class of interest. The degree of support for the hypothesis H_1 is measured with *test statistic*, T(x) which is a function of feature vector x, $x \in \mathbf{X}$. Under the $MVN[\mathbf{O}_q, \mathbf{I}_{qxq}]$ assumption of $f_x(x|C_{int})$, we will use the test statistic $T(x)=x^T x$. Now, the significance testing rejects a sample x if $T(x)=x^T x > \lambda$. The threshold λ specifies the *rejection region* in the feature space, thus controls the omission error which is denoted by ε_1,

$$\varepsilon_1 = P\{ \ T(x) > \lambda_a \mid H_1 \ \} \leq 1 - a, \ 0 \leq a \leq 1 \tag{1}$$

The value, $(1-a)$ defines the maximum allowable omission error and is called the *significance level* or *rejection* probability. In this paper, we call the parameter a the *acceptance probability* and use it in the derivations. The threshold associated with a, denoted by λ_a, can be obtained by solving,

$$\int_0^{\lambda_a} f_Y(y \mid c_{int}) dy = a \tag{2}$$

where $f_Y(y|C_{int})$ is the conditional distribution of $y=x^T x$, under the hypothesis H_1. (The notation of H_1 and C_{int} will be used interchangeably). When $y=x^T x$, $f_Y(y|C_{int})$ is known to be the chi-squared distribution with q degrees of freedom[9]. The commission error, denoted by ε_2, is generally very difficult to control, since its evaluation requires unavailable statistical knowledge about all alternatives. Increasing the acceptance probability a reduces the omission error at the rate of 1, but, at the same time, increases the commission error whose rate of increase is dependent on the closeness of the distribution of the others class to the class of interest. To avoid potentially excessive omission or commission errors, the significance level (equivalently, the acceptance probability) must be carefully determined by checking the relative distribution of data samples with respect to the class of interest.

Ⅲ. OPTIMAL SIGNIFICANCE TESTING

A. Omission and Commission Errors as Functions of Acceptance Probability α

One can compute the omission error, $\varepsilon_1(\alpha)$ in terms of the acceptance probability α by dividing the number of C_{int} samples rejected at an acceptance probability α with N_1 where N_1 is the number of samples belonging to C_{int} in the data set **X** and is unknown. Similarly, the commission error $\varepsilon_2(\alpha)$ is obtained by dividing the number of accepted C_{others} samples by $N_2 = N\text{-}N_1$.

$$\varepsilon_1(\alpha) = \frac{N_1 - \alpha \cdot N_1}{N_1} = 1 - \alpha \qquad (3.a)$$

$$\varepsilon_2(\alpha) = \frac{N(\alpha) - \alpha \cdot N_1}{N - N_1} \qquad (3.b)$$

$N(\alpha)$ is the expected number of data samples accepted with the acceptance probability α, written as,

$$N(\alpha) \equiv N \int_0^{\lambda_\alpha} f_Y(s)\,ds, \; 0 \le \alpha \le 1 \qquad (4)$$

where $f_Y(y)$ is the mixture probability density function of y, $y = x^T x$, $y \ge 0$, and λ_α is the threshold corresponding to acceptance probability α in eq.(2). Although the mixture density $f_Y(y)$ is not available a priori, note that it can be easily estimated using the y values where $y = x^T x$, $x \in \mathbf{X}$. Thus, one can easily estimate $N(\alpha)$. With respect to α, $\varepsilon_1(\alpha)$ is a strictly decreasing function with slope -1 and $\varepsilon_2(\alpha)$ is a monotonically increasing function, but, the actual rate of increase of $\varepsilon_2(\alpha)$ is dependent on the behavior of $N(\alpha)$. The evaluation of $\varepsilon_2(\alpha)$ generally requires N_1, or equivalently, the prior probability π_{int}.

B. Optimality Criteria

The optimal acceptance probability α is dependent on its optimality criterion. For example, α can be selected solely on the basis of the omission or the commission error, or, it can be selected based on a criterion which is basically a weighted sum of the two. In many situations, a simple average of them,

$$E_1(\alpha) \equiv \frac{1}{2}[\varepsilon_1(\alpha) + \varepsilon_2(\alpha)] \qquad (5.a)$$

serves as a good candidate for assessing optimality. From a classification point of view, this corresponds to minimizing the class-averaged classification error. On the contrary, the overall classification error corresponds to the Bayes total probability error criterion which minimizes,

$$E_2(\alpha) \equiv \pi_{int}\varepsilon_1(\alpha) + \pi_{others}\varepsilon_2(\alpha) \qquad (5.b)$$

the sum of two errors weighted with the prior probabilities. The weights in eq.(5.b) can be generalized by allowing different cost between omission and commission errors as,

$$E_3(\alpha) \equiv A \cdot \pi_{int}\varepsilon_1(\alpha) + \pi_{others}\varepsilon_2(\alpha) \qquad (5.c)$$

Constant A, where A > 0, is the cost on making the *omission* error relative to the cost of making the *commission* error being 1. The criteria in eq.(5.a,b) can be considered to be special cases of $E_3(\alpha)$ since $E_3(\alpha)$ with A=1 becomes $E_2(\alpha)$, and setting $A = \pi_{others}/\pi_{int}$ makes $E_3(\alpha)$ equivalent to $E_1(\alpha)$. In this sense, the criterion in eq.(5.c) is called the "generalized" total classification error criterion. In following discussions, only the criterion in eq.(5.c) will be used since each of the others can be derived as a special case of this criterion by setting an appropriate value of A.

The class-averaged classification error criterion is a very useful indicator of classification performance especially when there are large differences in prior probabilities since the overall classification accuracy is dominated by the performance of the classes having dominant prior probabilities. The class-averaged classification error in eq.(5.a) is desirable optimality criterion in applying significance testing to the partially supervised classification since the number of class-of-interest samples is in general much less than that of the others class.

C. Estimating Optimum Acceptance Probability

The optimal acceptance probability a can be obtained by minimizing $E_3(a)$ with respect to a over the interval, $0 \leq a \leq 1$. That is, by equating the first order derivative of $E_3(a)$ in eq.(6.a) to 0, and checking the sign of the second order derivative in eq.(6.b).

$$\frac{dE_3(a)}{da} = \frac{1}{N}\left[\frac{dN(a)}{da} - (1+A) \cdot N_1\right] = 0 \qquad (6.a)$$

$$\frac{d^2E_3(a)}{da^2} = \frac{1}{N}\frac{d^2N(a)}{da^2} \qquad (6.b)$$

Note that solving eq.(6.a) requires, in general, knowledge of N_1, or, equivalently, the prior probability π_{int}. Substituting the first order derivative of $N(a)$ in eq.(4) with respect to a into eq.(6.a) results in,

$$Nf_Y(\lambda_a) = (1+A)N_1 f_Y(\lambda_a \mid C_{int}) \qquad (7)$$

The first order derivative of $E_3(a)$, being always positive in $0 \leq a \leq 1$, indicates that $Nf_Y(\lambda_a)$ on the left side of eq.(7) is always larger than the right side, $(1+A)N_1f_Y(\lambda_a|C_{int})$ for all a in the interval [0,1]. Since (1+A) > 1, this means that the data samples expected to be in the infinitesimal region $(\lambda_a,\ \lambda_a + d\lambda_a)$ are always more than the expected number of C_{int} samples in the region; thus, considerable commission error will result no matter how restrictive the acceptance probability becomes. Therefore, the optimum value of a is expected to be 0. On the other hand, the first order derivative of $E_3(a)$, being always negative in the closed interval, indicates, by the same token, that the data points expected to be in the infinitesimal region $(\lambda_a,\ \lambda_a + d\lambda_a)$ are always less than the expected number of C_{int} samples in the region (which is weighted by (1+A)), therefore, the possibility of commission is very low. This will allow acceptance probability a to increase up to 1. Other than these two extremes, the minimum point of $E_3(a)$ will be located where the degree of increase of the weighted commission error starts to surpass the decrease of the weighted omission error. The prior probabilities and relative cost A determine the actual balancing between omission and commission errors. Due to the closed interval of a, a minimum of $E_3(a)$ always exists and so does an optimum a, even if there may be no solution satisfying eq.(6.a) and the positiveness of eq.(6.b). Suppose solutions satisfying these two conditions do exist, and denote a set of those solutions as S.

$$S \equiv \{a \mid \frac{dE_3(a)}{da} = 0 \text{ and } \frac{d^2E_3(a)}{da^2} > 0,\ 0 \leq a \leq 1\}$$

The elements in S correspond to the (local) minima of $E_3(a)$. The global minimum can be selected by comparing the actual values of $E_3(a)$ at different a's in S in the following way : suppose a_i, a_j are elements in S, then, the difference, $E_3(a_i)$ - $E_3(a_j)$ is written as,

$$E_3(a_i) - E_3(a_j) = \frac{\triangle_{ij}}{N} \qquad (8)$$

$$where,\ \triangle_{ij} \equiv [N(a_i) - N(a_j) - (a_i - a_j) \cdot (1+A) \cdot N_1]$$

By checking the signs of the $\triangle_{ij}$'s, the acceptance probability which achieves the global minimum of $E_3(a)$ can be selected from the set S. Notice that solving eq.(7) and evaluating eq.(8) requires N_1, but, under the class-averaged classification error criterion of $E_1(a)$, it can be evaluated even without knowing N_1 since substituting $A = \pi_{others}/\pi_{int} = N_2/N_1$ gives $(1+A)N_1 = (1+N_2/N_1)N_1 = N$, independent of N_1. This property of the class-averaged classification error criterion is very useful in actual application of this method, since the number N_1 is unknown in most problems. Note that the class-averaged classification error criterion of $E_1(a)$ makes very good sense in applying significance testing to the partially supervised classification problem because the classification performance is not dominated by the relatively large prior probability of the others class.

Ⅳ. Experiments and Discussion

To test the performance of the proposed method and application to the partially supervised classification problem, experiments are carried out with simulated Gaussian data.

For the class of interest, 1000 bivariate Gaussian samples are generated with zero mean and an identity covariance matrix. For the others class, 2000 bivariate Gaussian samples are generated with mean $[d, 0]^T$, $d > 0$, and an identity covariance matrix. With this set-up, the exact amount of overlap when the distance between two class means is d can be calculated as,

$$Overlap(d) = 1 - \frac{2}{\sqrt{2\pi}} \int_0^{d/2} exp(-\frac{1}{2}s^2)ds$$

The term "overlap" is defined as the volume shared by the two probability density functions. By varying d from 0.1 to 5 in steps of 0.1, data sets with different degrees of overlap can be simulated: d=0.1 simulates 96.02% of overlap, and d=5 produces only 1.24% of overlap. To avoid any random error due to the data generation process and its effect on evaluating the experimental result, data sets are generated 50 times with different seed numbers, and their average is used in comparison. Since the class-of-interest data are Gaussian with zero mean and identity covariance matrix, the test statistic $y=x^Tx$ is used.

At first, various different acceptance probability α's from 0.01 to 0.99 in steps of 0.01 are tested to see its effect on classification accuracy of a partially supervised supervised classifier using significance testing. As expected, the omission error decreases linearly with respect to the acceptance probability with slope = -1, and the slope of the commission error increase depends on the degree of overlap between the two distributions. When d = 0.5 (~ 80.26% of overlap), the commission error increases almost linearly with respect to a. This is due to the substantial closeness of the two distributions. When there is effectively no overlap such as in the case d = 4.5 (2.44% of overlap), the commission error is observed to stay very low, virtually insensitive to α.

To evaluate the accuracy of estimated α values, (true) optimal acceptance probabilities are manually determined by changing α from 0.01 to 0.99 in steps of 0.01 under the selected optimality criterion. These manually determined values are denoted by "scanned," and used as references in evaluating the accuracy of the estimates obtained by the proposed method. The estimated acceptance probabilities with both the class-averaged and the total classification error criteria are shown in Fig. 1. When applying the total classification error criterion, the true values of prior probabilities are used.

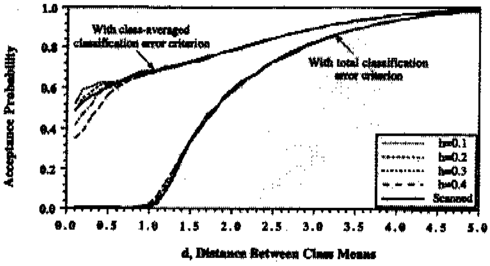


Fig. 1 Estimated optimal acceptance probability α versus d, the distance between two class means. Solid lines show the manually selected acceptance probabilities. Various dotted lines show the estimated optimal acceptance probabilities using the proposed method with different Parzen window sizes.

The density estimate required for N(α) is obtained by employing a Gaussian Kernel-based Parzen density estimate with the data set, augmented by positive reflection [12]. Even though an appropriate kernel window size h is computed as 0.2 based on [11], several different values are also tested to see its effect on the estimated acceptance probabilities. Figure 1 shows that the estimated α values follow very closely those manually determined true values especially when the distance d is large. The optimal acceptance probability based on the total classification error criterion is near 0 when d is not large enough, since the total classification error is an increasing function of acceptance probability for those small d values. For example, when d < 1.0, the commission error increases almost at the same rate as the omission error decreases due to the significant overlap between the two class distributions. Because the prior probability of C_{int} is less than that of C_{others}, the omission error is weighted less than the commission error under the total classification error criterion. This explains why the acceptance probabilities for d < 1.0 are almost zero under the total classification error criterion.

Figure 1 also shows some degree of difference between the estimated and the manually determined acceptance probability under the class-averaged classification error criterion when d < 1.0; in this region, the actual curve of class-averaged classification error is observed nearly flat with α value in the range 0.45～0.5; therefore, an exact location of the minimum of the class-averaged classification error is expected to be hard to pinpoint. It is also sustained by an experimental observation that there is a relatively large standard deviation not only in the estimated but also in the manually selected optimum α values. The same argument can be made for the deviations in the region 1.0 < d < 2.0 under the total classification error criterion.

The result of classification errors is given in Fig.2. In spite of those discrepancies in estimated α values, there is not much difference in the resulting class-averaged and total classification errors. Since less than 1% of the differences are observed with varying Parzen window sizes under both optimality error criteria, Fig.2 shows only the classification results with h=0.2. Note that the classification based on significance testing deals with only the one-dimensional values of the selected test statistic T(x), therefore, the dimensionality reduction of feature vectors to one-dimensional space causes information loss in classification [2][3]. To see its effect, a (fully) supervised maximum likelihood classifier (denoted as "REL-ML") and a maximum a posterior classifier (denoted as "REL-MAP")[2][4] are designed in the original q-dimensional space with known class statistics of C_{int} and C_{others}, and their performances are also drawn in the Fig.2.

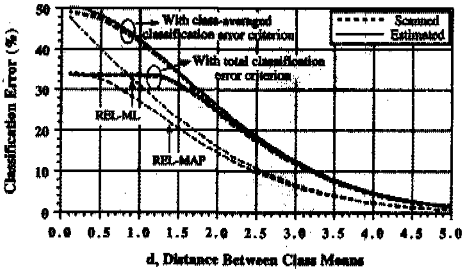


Fig. 2 Class-averaged and total classification error versus the distance between the two class means; Acceptance probabilities are estimated with the class-averaged and total classification error criteria. "REL-ML" and "REL-MAP" are respectively results of the fully supervised maximum likelihood and the maximum a posterior classifier. Parzen window size h = 0.2.

As seen in Fig.2, the estimated optimal acceptance probabilities result in almost the same performances with manually determined values under both optimality conditions. By the way, a maximum of about 12% error increase due to the dimensionality reduction is observed. This is what the simplicity of a classifier based on the significance testing has to pay for.

Density estimation without reflected data[12] is expected to introduce under-estimation of the probability density $f_Y(y)$ near y = 0 due to using a symmetric kernel function with only positive y values. This under-estimation in $f_Y(y)$ and subsequently in N(α) near y=0 would cause under-estimation of commission errors, therefore, the optimal acceptance probability estimates are expected to be larger than they should be. Since the Gaussian kernel function rapidly decreases as its argument becomes larger, the effect of under-estimation due to lack of reflection would exist only in the region near y=0. Experimentally, no difference is observed under the class-averaged classification error criterion; this is because the optimum a is much larger than 0 as seen in Fig.2. However, in the case of the total classification error criterion, the estimated acceptance probabilities without data reflection are observed to be larger by as much as 0.2 compared to those with data reflection in the region of d < 1.5. But, no differences are seen when d > 1.5. Greater difference is noticed as the window size h becomes larger. This is because a large window size has more reflected samples in the summation of the kernel function values. The reflection technique in estimating a probability density function of $y=x^Tx$ is observed to be necessary if the acceptance probabilities are expected to be near zero. The discrepancies in acceptance probabilities caused by not using data reflection are observed to result in as much as 5% difference of the total classification error in the region d < 1.5.

V. Conclusion

In this paper, the problem of estimating the optimal acceptance probability (equivalently, significance level) is addressed in the context of applying significance testing to partially supervised classification. As the optimality criteria, both class-averaged and generalized total classification error criteria are considered. It is shown that if the class of interest does not consist of multiple sub-classes, the optimum acceptance probability under the class-averaged classification error criterion can be quite accurately estimated without any prior knowledge except the probability density function of the class of interest.

This estimation method for acceptance probability should be very useful when one does not have enough prior knowledge about the data set to select the proper acceptance probability. This unsupervised estimation procedure can replace the lengthy and tedious process of manual selection of acceptance probability especially when the given class of interest consists of a large number of sub-classes.

References

[1] P. H. Swain and S. Davis (editors), "Fundamentals of Pattern Recognition in Remote Sensing," *Remote Sensing - The Quantitative Approach*, McGraw-Hill Book Company, New York, 1978.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd edition, Academic Press, New York, 1990.

[3] B. Jeon and D. A. Landgrebe, "Partially Supervised Classification Using Weighted Unsupervised Clustering," IEEE Trans. Geoscience and Remote Sensing, vol. 37, No. 2, March 1998.

[4] J. A. Richards, *Remote Sensing Digital Image Analysis, An Introduction*, 2nd Edition, Spring-Verlag, 1993.

[5] K. Fukunaga, R. R. Hayes, and L. M. Novak, "The Acquisition Probability for a Minimum Distance One-class Classifier," IEEE Trans. Aerospace and Electronic Systems, AES-23, pp. 493-499, 1987.

[6] C. W. Therrien, T. F. Quatieri, and D. E. Dudgeon, "Statistical Model-Based Algorithms for Image Analysis," Proceedings of IEEE, Vol. 74, No. 4, pp. 532-551, April 1986.

[7] T. F. Quatieri, "Object Detection by Two-dimensional Linear Prediction," Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, ICASSP-83, pp. 108-111, Apr. 1983.

[8] M. G. Bello, "A Random-Field Model-Based Algorithm for Anomalous Complex Image Pixel Detection," IEEE Trans. Image Processing, Vol. 1, No. 2, pp. 186-196, April 1992.

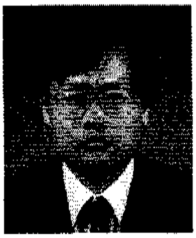
[9] A. Drake, *Fundamentals of Applied Probability Theory*, McGraw-Hill, New York, 1967.

[10] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*, John Wiley & Sons, New York, 1968.

[11] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.

[12] L. I. Boneva, D. G. Kendall and I. Stefanov, "Spline Transformations: Three New Diagnostic Aids for the Statistical Data-analysis (with Discussions)," Journal Royal Statist. Soc. B, 33, pp. 1-70, 1971.

전 병 우(Byeungwoo Jeon) 정회원



1985년 : 서울대학교 전자공학과 학사

1987년 : 서울대학교 전자 공학과 석사

1992년 : Purdue Univ, School of Elec. Eng. 박사

1993년~1997년 8월 : 삼성전자 멀티미디어 연구소 수석연구원

1997년 9월~현재 : 성균관대학교 전기전자 컴퓨터 공학부 조교수

<주관심 분야> 멀티미디어, 영상압축, 영상인식

David A. Landgrebe

Dr. Landgrebe holds the BSEE, MSEE, and PhD degrees from Purdue University. He is presently Professor of Electrical and Computer Engineering at Purdue University. His area of specialty in research is communication science and signal processing, especially as applied to Earth observational remote sensing.