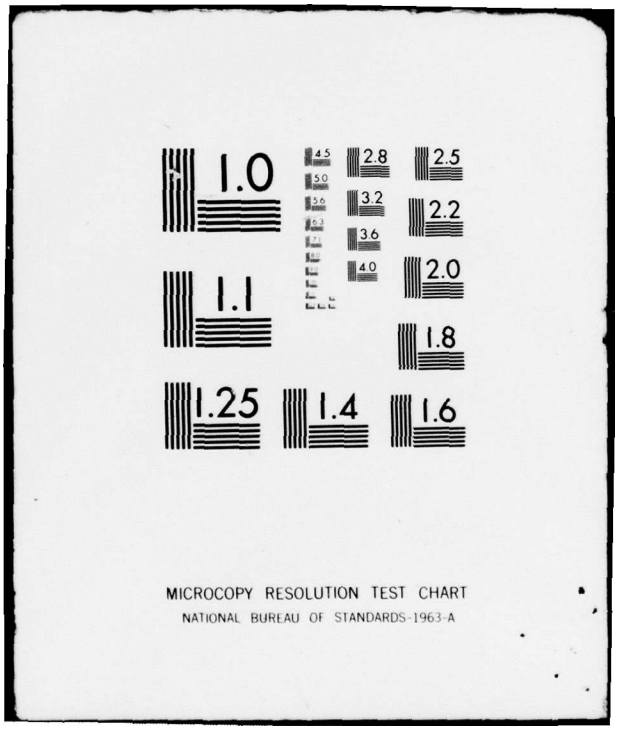
AD-A055 745 SOUTHEASTERN MASSACHUSETTS UNIV NORTH DARTMOUTH DEPT --ETC F/G 12/1
A REVIEW OF STATISTICAL PATTERN RECOGNITION.(U)
MAY 78 C H CHEN AFOSR-76-2951

UNCLASSIFIED EE-78-2 AFOSR-TR-78-1040 NL

| OF |
AD
A055745

END
DATE
FILMED
8 -78
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(2)

TECHNICAL REPORT SERIES IN INFORMATION SCIENCES
(Dr. C. H. Chen, Principal Investigator)

SOUTHEASTERN MASSACHUSETTS UNIVERSITY

ELECTRICAL ENGINEERING DEPARTMENT

NORTH DARTMOUTH, MASS. 02747 U.S.A.

Grant AFOSR 76-2951
TR No. EE-78-2
May 19, 1978

A REVIEW OF STATISTICAL PATTERN RECOGNITION

by

C. H. Chen
Department of Electrical Engineering
Southeastern Massachusetts University
North Dartmouth, Massachusetts 02747

Abstract

This paper carefully examines the current status of the statistical pattern recognition by the topics: classification rules, feature extraction, contextual analysis, etc. Important but unsolved problem areas are also explored. The relationship between the statistical pattern recognition and signal processing is also considered.

A Review of Statistical Pattern Recognition

C. H. Chen

I. Introduction

After more than twenty years of progress, the theory and applications of statistical pattern recognition are now well developed. A number of textbooks [1-11] have been available. The limitations of the statistical pattern recognition are also evident: the patterns are not characterized by the statistical information alone and many useful statistical properties cannot be fully exploited with available mathematical statistics. Like many other fields there is a wide gap between theory and practice. The limitation of the finite sample size is mainly responsible for such a gap. The finite sample size effect is the one among ten problem areas [12] in statistical pattern recognition for which the solutions are much needed.

In this paper the current status of the statistical pattern recognition is reviewed by topics including classification rules, feature extraction, contextual analysis, supervised and unsupervised learning and clustering, finite sample size effects, and computational recognition complexity. Other important but unsolved problem areas are examined. The relationship between the statistical pattern recognition and signal processing is also considered.

II. The Classification Rules

Statistical pattern recognition makes use of the decision theoretic approach to pattern recognition. The fundamental assumption is that the pattern are random in nature and thus can be described statistically in parametric or nonparametric forms. The recognition problem essentially consists of preprocessing, feature extraction and selection, and classification (decision making) along with training or learning process. A good classification is almost always the main objective of a recognition system. Two most well known statistical classification rules are the Bayes decision rule and the nearest-neighbor decision rule.

Let x be a vector measurement of a pattern sample, and m be the number of classes. The Bayes decision rule minimizes the average risk with respect to the given a priori probabilities P_i, $i = 1,2,\ldots,m$. For equal loss functions, the Bayes decision rule reduces to the maximum likelihood decision rule (MLDR) which chooses the class that maximizes the function

$$P_i\ p(x/\omega_i);\quad i = 1,2,\ldots,m$$

where the conditional probability densities $p(x/\omega_i)$ must be known or estimated. The optimal property of the Bayes decision rule is not always realized in practice because the required a priori knowledge is either unavailable or inaccurate. For two multivariate Gaussian densities with mean μ_i and covariance $\sum_i$, $i = 1,2$, the MLDR is to assign x to the class for which

$$(x - \mu_i)' \sum_i^{-1} (x - \mu_i) - \ell n(P_i^2/|\sum_i|) \qquad (1)$$

is the minimum. It is not unusual to find in practice [13] that a modified MLDR which chooses the minimum of the form,

$$(x - \mu_i)' \sum_i^{-1} (x - \mu_i) \qquad (2)$$

can perform better than the MLDR. This is an example of the gap between theory and practice. The performance of the Bayes decision rule or the Bayes error probability in general cannot be expressed with a closed form. The error estimate which critically edpends on the sample size is by itself an fundamental problem in statistics (see e.g. [14])

The nearest neighbor decision rule (NNDR) identifies the vector sample x with the class of its nearest neighbor; nearness being measured by the Euclidean distance. For k-NNDR, the decision is based on the majority vote of k nearest neighbors. The advantage of the NNDR is that its asymptotic error rate is upper bounded by twice of the Bayes error. The NNDR is nonparametric because the information on probability densities is not needed. An obvious drawback of the NNDR is that an extensive amount of distance computation is required. Procedures to reduce the computation

include the condensed NNDR, edited NNDR, selection of training samples, and the use of branch and bound algorithms. Other modifications of the NNDR include the distance weighted NNDR which can provide better recognition result in practice than the unweighted NNDR discussed above. Replacement of the Euclidean distance by the quadratic form given by Eq. (2) also demonstrated a superior recognition performance in practice [15].. The performance of the NNDR at small sample size is not clear as the limited available theoretical results are inconclusive. For moderate to large sample size, the NNDR performance is comparable to the MLDR.

The reject option has been considered for both Bayes decision rule and the NNDR. The errors can be reduced at the expense of some rejects. The error-reject trade-off is an additional consideration in the reject option (see [16] for recent result).

Linear, piecewise linear, and quadratic discriminant functions have been extensively investigated especially in the statistical literatures. However, the closed form error probability expressions are generally unavailable except in the simple case of multivariate Gaussian densities with unequal mean and equal covariance matrices. The use of the MLDR is implied for the parametric discriminant analysis and the optimization criterion is the minimum error probability. The Fisher's linear discriminant is a nonparametric technique that maximizes the ratio of between-class scatter to within-class scatter in the one-dimensional space on which the vector measurements are projected. This projection is a many-to-one mapping and in theory cannot possibly reduce the minimum attainable error probability.

For complex patterns such as images, a multi-stage decision-tree classifier has been shown experimentally to have a better overall performance than the conventional single-stage classifier [17][18]. However, the classification time increases due to the complexity of computation. A linear binary tree classifier

can be used [19] to take advantage of the accuracy of a decision-tree classifier and to use the linear discriminant function at decision stages to reduce the classification time. With pre-designed tree structure, the overall computation time can be less than ten percent of that of a single-stage classifier. Although different feature subset may be used at each decision stage, the search for an otpimum feature subset requires additional computation. The problem of optimizing the decision tree structure has been considered (see e.g. [20]). The methods of reducing the computational complexity considered include clustering the decision rules, and the use of branch and bound procedure to find efficient decision rule and for feature assignment, etc. The decision tree classifier is the most promising classification mechanism for increasingly complex recognition problems in the future. Features can be mathematical, structural or various combinations.

Although the sequential decision procedure is, theoretically speaking, suitable mainly for independent identically distributed measurements, the flexibility allowed by feature ordering or even on-line feature ordering is the most attractive capability of the sequential decision procedure.

The table look-up decision rule stores the decision rule itself rather than the densities. The vector measurement x is used as an address to a table which look-up the class assignment for x. The table which is stored in the memory assigns a class to each (quantized) vector in the measurement space. Procedures to reduce the memory requirements and to speed-up the decision assignment time have been considered ([21][22]).

Other generalization of the conventional decision theory framework is the simultaneous membership of a measurement in several classes which has the origin of "degree of membership" from fuzzy set theory. The compound decision rules and the finite sample size effects in sample-based classification rules will be discussed in later sections.

III. Feature Extraction

The mathematical features as well as the structural features are best suited for automatic recognition although they may not necessarily have physical meaning or may be quite different from features derived by human recognition process. A fundamental approach to extract features in statistical pattern recognition is by evaluating a number of available features to select a small subset of good features. Such evaluation can be based on the direct estimate of the error probability. Many feature selection criteria have been proposed for feature evaluation including various distance and information measures (see e.g. [23][24]). These measurements are very effective even though they do not always choose the feature set that has the smallest error. The relative effectiveness of various measures has been considered [25]. These measures are also very useful for error estimates [26].

Another useful approach is the linear transformation methods. If a pattern can be completely described by the second order statistics, the Karhunen-Loeve transform is optimal in the mean square error sense. In addition to the fact that the second order statistics is not adquate for most patterns, the transform also requires excessive computation. It is a misconception that feature extraction is nothing more than dimensionality reduction and that the Karhunen-Loeve transform solves all mathematical feature extraction problems.

A realistic solution to feature extraction must take into consideration the nature of patterns, the a priori knowledge available, and the specific requirements and constraints of the given recognition task. Although exhaustic search is about the only way to find the best feature set, efficient feature set search procedures are most needed [27] to provide a computationally feasible solution.

Feature extraction and selection is important not only for pattern recognition but also useful to signal processing and communications. Properly selected

feature subset can represent a compression of the original signal so that the transmission requirement such as bandwidth can be greatly reduced. However, feature selection differs from signal selection in communications in one important aspect; the additive white noise usually does not apply to the pattern recognition problem. To extract the right features that truly characterize a pattern is a real challenge to human intelligence. Although much has been studied, feature extraction will remain to be a key problem in pattern recognition.

IV. Contextual Analysis

A major weakness of statistical pattern recognition is the difficulty to take the contextual relations into account in the recognition process. The compound decision theory appears to be the closest statistical theory that can take the contextual information into account. When a statistical decision problem is repeated n times, with no relationships among the individual problems, the compound decision rule makes use of the information from all measurements from the n repetitions to make decisions on individual problems. In character recognition of a text, for example, decisions have to be made on individual characters. The contextual information in terms of transition probabilities among characters can be utilized to improve the recognition for individual characters. Similarly in image recognition, individual picture elements or subimages may have to be classified. The information on the correlation among picture elements or subimages should be used for better classification. Although very little theoretical result is available to measure the amount of performance improvement due to the use of contextual information, experimental results have all demonstrated the available improvement. To implement the compound decision rule, Markov chain, model of stationary stochastic process for the pattern, and coding of spatial correlation parameters [28] are among the useful tools.

Consider the recognition of each subimage of an image. By assuming dependence only on four adjacent subimages, the compound decision rule is to choose the class

which maximizes

$$p(x_o/\omega_k)P(\omega_k) \prod_{j=1}^{4} p(x_j/\omega_k) \qquad (3)$$

where $\omega_k = 1,2,\ldots,m$ and x_o is the vector measurement of the subimage under consideration. If we assume the dependence on all eight neighboring subimages, then the expression inside the product sign should have the conditional probability densities of all eight neighbors. Experimental result has demonstrated [29] that there is very little performance difference between four and eight neighbors.

While there is very much to be done in image recognition using the contextual information to classify a whole image or individual subimages (or picture elements), there has been very significant progress in the character recognition area (see e.g. [30][31]).

V. Supervised and Unsupervised Learning and Clustering

Learning is needed in pattern recognition to establish the required statistical knowledge, from samples, such as the statistical parameters, probability densities, or even the decision boundaries. When the samples are of known classification, learning is supervised; otherwise it is unsupervised. In terms of the statistical framework, the supervised learning follows exactly the classical Bayesian and maximum likelihood estimation theories.The mixture estimation and decomposition in statistics is one approach to unsupervised learning. Much details on the learning algorithms as well as the decision-directed learning are available in pattern recognition texts [1-11]. It is important to note that the criterion of minimizing the mean-square error between the estimated and true parameters is used almost exclusively in learning and estimation. While the objective of classification is the minimum error probability, there is no guarantee that the learning algorithms will result in minimum classification error. Some effort has been made to design learning algorithms using window functions to minimize directly the classification error [32]. However the convergence rate may be slow. In addition

to properly selecting the window parameter, other procedures should be examined to speed up the convergence. A good understanding of the relationship [33] between estimation and decision is necessary. More flexible structures for the learning process should be considered. For example, the initial learning phase may be the conventional minimum mean-square error criterion. The subsequent learning phase can be based on the minimum error probability criterion. Another example is that a supervised learning process can be switched to unsupervised learning or vice versa. Of course the optimum usage of each learning phase would be a new problem to be examined [34].

Clustering is an important subject by itself in statistical data analysis, although it may be considered as unsupervised learning in pattern recognition. Clustering can be defined as a partition of the set of vector measurements such that each measurement will be assigned to one and only one set among a collection of disjoint sets. A recent discussion on the subject is in [35], in addition to the texts [1-11]. The problem of clustering individuals can be considered within the context of a mixture of distributions [36]. Discussion of the cluster validity problem is in [37].

VI. Finite Sample Size Effects

In practical recognition problems the sample size is limited. The actual recognition performance may be quite different from that theoretically predicted based on infinite sample size. Indeed the finite sample size and its associated dimensionality problem is fundamental to all pattern recognition problems. For example, the decision rules in practice are sample-based. Expected errors of the sample-based classification rules generally do not have closed form solution at small sample size. Distance and information measures evaluated under finite sample size may be highly inaccurate. A general discussion of the finite learning sample size problem is in [38][39][40] among others.

The best way to reduce the finite sample size effect is to increase the sample size with respect to the dimensionality. For images the dimensionality includes the numbers of picture elements and the quantization levels. The relationships among the performance, sample size, and dimensionality are highly nonlinear. In general when the sample size is moderately large to large, the effects of finite sample size are not very significant. A thorough study of the subject is much needed as it will certainly be helpful to design a reliable recognition system for a given set of features.

VII. Computational Recognition Complexity

The term "computation complexity" has a different meaning at different situations and is not well defined for pattern recognition researchers. The Kolmogorov information-theoretic computational complexity is defined as the minimum length of the program to obtain an object from data. While in linear discrimination the complexity of the classifier is usually identified with the dimensionality of the vector measurement, the discriminating capability of Boolean classifiers is determined not only by dimensionality of the feature vectors but also by the type of combinations these features are permitted to undergo. In this case we talk about the combinational complexity of the decision rule. Intuitively the complexity concept can give us a feeling of what is complex and what is less complex. So the complexity should be a relative not an absolute measure. A more familiar complexity definition to engineers is the amount of computational effort including time and cost to accomplish a recognition task. To be machine independent, the complexity will include mainly the number of manipulations such as the multiplication and comparison operations. The recognition complexity based on this definition can be reduced by proper implementation techniques such as the use of sequential-parallel operations, etc.

For the overall recognition complexity of a recognition system, the trade-off between feature extraction and classification must be considered. A complicated feature extraction process results in a few but good features. The resulting classifier can be a very simple one. If no feature extraction effort is made so that a large number of features are used, the required classification and learning process will be very complicated. The problem of determining an optimum overall recognition time has not been considered. The solution to this problem should be particularly useful for realtime pattern recognition.

VIII. Other Problem Areas

In addition to the topics considered above, there are a number of other problem areas where the solutions are partially available or completely unavailable.

1. Learning and classification of nonstationary patterns. Only special cases were examined.
2. A truly optimal recognition system that optimizes jointly the preprocessing, feature extraction, and classification and learning. Solution is not available.
3. Statistical and syntactic mixed model. Much has been said but little success is reported.
4. Automatic generation of recognition rules. No solution is available.
5. Interactive pattern recognition. A very significant progress has been made to provide man-machine interaction in pattern recognition.

IX. Relationships with Signal Processing

Many statistical pattern recognition techniques such as feature extraction and classification can be considered as "nonlinear" signal processing. On the other hand many digital signal processing techniques are especially needed for the preprocessing phase of the recognition process. However, in signal processing the emphasis is on manipulation of patterns of a single class while in pattern recognition the emphasis is on the difference among the patterns from several classes. Integration of processing and recognition into one system has been necessary in many applications.

References

(Author's note: No attempt is made to provide an exhaustic list of all contributions in the statistical pattern recognition. The author offers his apology for not listing the important work of many individuals. The ASI refers to the NATO Advanced Study Institute on Pattern Recognition and Signal Processing, Paris, June 1978.)

1. G.S. Sebestyen, "Decision Making Processes in Pattern Recognition", Macmillan, 1962.

2. K.S. Fu, "Sequential Methods in Pattern Recognition and Machine Learning", Academic Press, 1968.

3. H.C. Andrews, "Introduction to Mathematical Techniques in Pattern Recognition", Wiley, 1972.

4. K. Fukunaga, "Introduction to Statistical Pattern Recognition", Academic Press, 1972.

5. W. Meisel, "Computer-Oriented Approaches to Pattern Recognition", Acaemic Press, 1972.

6. E.A. Patrick, "Fundamentals of Pattern Recognition", Prentice-Hall, 1972.

7. R.O. Duda and P.E. Hart, "Pattern Classification and Science Analysis", Wiley, 1973.

8. C.H. Chen, "Statistical Pattern Recognition", Hayden Book Company, 1973.

9. T.Y. Young and T.W. Calvert, "Classification Estimation, and Pattern Recognition", American Elsevier, 1973.

10. J.R. Ullmann, Pattern Recognition Techniques", Crane, Russak & Co., 1973.

11. J.T. Tou and R.C. Gonzales, "Pattern Recognition Principles", Addison-Wesley Publishing Co., 1974.

12. C.H. Chen, "Statistical pattern recognition: review and outlook", TR EE-75-4, June 1975.

13. J.K. Chang, "Modified maximum likelihood decision rule and minimax Bayes decision rule", Proc. of the Third International Joint Conference on Pattern Recognition, November 1976.

14. G.T. Toussaint, "Bibliography on estimation of misclassification", IEEE Trans. on Information Theory, Vol. IT-20, pp. 472-479, July 1974.

15. C.H. Chen, "Seismic pattern recognition", Geoexploration Journal, April 1978.

16. P.A. Devijver, "Error-reject relationships in nearest neighbor decision rules", Proc. of the Third International Joint Conference on Pattern Recognition, November 1976.

17. C.L. Wu, "The decision tree approach to classification", Ph.D. thesis, School of Electrical Engineering, Purdue University, May 1975.

18. H. Hauska and P.H. Swain, "The decision tree classifier: design and potential", Proc. of Symposium on Machine Processing of Remotely Sensed Data, June 1975.

19. K.C. You and K.S. Fu, "An approach to the design of a linear binary tree classifier", Proc. of Symposium on Machine Processing of Remotely Sensed Data, June 1976.

20. A.V. Kulkarni and L.N. Kanal, "An optimiation approach to hierarchical classifier design", Proc. of the Third International Joint Conference on Pattern Recognition, November 1976.

21. W.G. Eppler, "An improved version of the table look-up algorithm for pattern recognition", Proc. of the Ninth International Symposium on Remote Sensing of Environment, April 1974.

22. R.M. Haralick, "The table look-up rule", Proc. of the Third International Joint Conference on Pattern Recognition, November 1976.

23. L.N. Kanal, "Patterns in pattern recognition", IEEE Trans. on Information Theory, Vol. IT-20, pp. 697-722.

24. C.H. Chen, "On information and distance measures, error bounds and feature selection", Information Sciences Journal, Vol. 10, pp. 159-173, 1976.

25. C. H. Chen, "Theoretical comparison of a class of feature selection criteria in pattern recognition", IEEE Trans. on Computers, Vol. C-20, pp. 1054-1056, September 1971.

26. P.A. Devijver, "On a new class of bounds on Bayes risk in multihypothesis pattern recognition", IEEE Trans. on Computers, Vol. C-23, pp. 70-80, January 1974.

27. J. Kittler, "Feature set search algorithms" in this Proceedings.

28. T.S. Yu and K.S. Fu, "Statistical pattern recognition using contextual information", TR-EE-78-17, Purdue University, 1978.

29. J.R. Welch and K.G. Salter, "A context algorithm for pattern recognition and image interpretation", IEEE Trans. on Systems, Man and Cybernetics, SMC-1, pp. 24-30, January 1971.

30. G.T. Toussaint, "The use of context in pattern recognition", Proc. of Pattern Recognition and Image Processing Conference, June 1977.

31. C.Y. Suen, "Handprinting recognition and standardization", presented at this ASI Conference.

32. H. Dotu, Learning algorithms for discriminant function solution of the minimum error classification problem", Proc. of the Third International Joint Conference on Pattern Recognition, November 1976.

33. W. Hodgkiss and L.W. Nolte, "On relationships between detection and estimation theory", this proceedings.

34. D.B. Cooper, "When should a learning machine ask for help", IEEE Trans. on Information Theory, Vol. IT-20, pp. 455-471, July 1974.

35. E. Diday and J.C. Simon, "Clustering analysis" in "Digital Pattern Recognition" edited by K.S. Fu, Springer-Verlag, 1976.

36. S.L. Sclove, "Population mixture models and clustering algorithms", Communications in Statistics, Vol. A6, pp. 417-434, 1977.

37. A.K. Jain, "Cluster validity", presented at this ASI

38. C.H. Chen, "Finite sample considerations in statistical pattern recognition", Proc. of Pattern Recognition and Image Processing Conference, May 1978.

39. L.F. Pau, "On finite learning sample size problems in pattern recognition", this proceedings.

40. S. Raudys and V. Pikelis, "On dimensionality, sample size and classification error in discriminant analysis", submitted for publication, 1978.

(14) EE-78-2

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
(19) (18) 1. REPORT NUMBER AFOSR-TR-78-1040	2. GOVT ACCESSION NO. (11) 3. RECIPIENT'S CATALOG NUMBER 19 May 78
4. TITLE (and Subtitle) (6) A REVIEW OF STATISTICAL PATTERN RECOGNITION.	5. TYPE OF REPORT & PERIOD COVERED (9) Interim rept.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) C.H./Chen	8. CONTRACT OR GRANT NUMBER(s) (15) ✓ AFOSR-76-2951
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Electrical Engineering Southeastern Massachusetts University North Dartmouth, Massachusetts 02747	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS (16) 61102F 2304/A2 (17) A2
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB, Washington, DC 20332	12. REPORT DATE May 19, 1978
	13. NUMBER OF PAGES 14 (12) 16p.
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS (of this report) UNCLASSIFIED
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper carefully examines the current status of the statistical pattern recognition by the topics: classification rules, feature extraction, contextual analysis, etc. Important but unsolved problem areas are also explored. The relationship between the statistical pattern recognition and signal processing is also considered.

407 932 MW

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)