LARS Technical Report 030178

# Bayesian Classification in a Time-varying Environment

Philip H. Swain

The Laboratory for Applications of Remote Sensing
Purdue University West Lafayette, Indiana
1978

| 1. Report No. | 2. Government Accession No. | 3. Recipient's Catalog No. |
|---|---|---|
| | | |

| 4. Title and Subtitle | 5. Report Date |
|---|---|
| Bayesian Classification in a Time-Varying Environment | March 1, 1978 |
| | 6. Performing Organization Code |

| 7. Author(s) | 8. Performing Organization Report No. |
|---|---|
| Philip H. Swain | 030178 |
| | 10. Work Unit No. |

| 9. Performing Organization Name and Address | 11. Contract or Grant No. |
|---|---|
| Laboratory for Applications of Remote Sensing<br>1220 Potter Drive<br>West Lafayette, Indiana 47906 | NAS9-14970 |
| | 13. Type of Report and Period Covered |

| 12. Sponsoring Agency Name and Address | 14. Sponsoring Agency Code |
|---|---|
| NASA | |

15. Supplementary Notes


16. Abstract

This paper deals with the problem of classifying a pattern based on multiple observation made in a time-varying environment. The identity of the pattern may itself change. A Bayesian solution is derived, after which the conditions of the physical situation are invoked to produce a "Cascade" classifier model. Experimental results based on remote sensing data demonstrate the effectiveness of the classifier.



| 17. Key Words (Suggested by Author(s)) | 18. Distribution Statement |
|---|---|
| pattern classification, multitemporal observations, remote sensing. | |

| 19. Security Classif. (of this report) | 20. Security Classif. (of this page) | 21. No. of Pages | 22. Price* |
|---|---|---|---|
| | | 18 | |


*For sale by the National Technical Information Service, Springfield, Virginia 22161

LARS Information Note 030178

# BAYESIAN CLASSIFICATION IN A TIME-VARYING ENVIRONMENT

Philip H. Swain

## Abstract

This paper deals with the problem of classifying a pattern based on multiple observations made in a time-varying environment. The identity of the pattern may itself change. A Bayesian solution is derived, after which the conditions of the physical situation are invoked to produce a "Cascade" classifier model. Experimental results based on remote sensing data demonstrate the effectiveness of the classifier.

Key words: pattern classification, multitemporal observations, remote sensing

# BAYESIAN CLASSIFICATION IN A TIME-VARYING ENVIRONMENT

Philip H. Swain[1]

## Introduction

We pose the following pattern classification problem:

A series of observations is made on a pattern in a time-varying environment. The identity of the pattern itself may change. It is desired to classify the pattern after the current observation is made, drawing on information derived from earlier observations plus knowledge about the statistical behavior of the environment.

An example of such a situation arises in remote sensing applications in which the sensor system can make multiple passes over the same ground area [1]. The identity of the ground cover may change between passes. In general it is desired to determine the current identity of the ground cover, but past observations can be helpful in accomplishing the identification.

## Approach

The classification strategy we shall develop is a Bayes optimal (minimum risk) strategy [2]. In the ordinary single

---

[1]Philip H. Swain is with the School of Electrical Engineering and the Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN 47907.

observation case, the approach is to select a decision rule so as to minimize the conditional average loss

$$L_X(\omega_i) = \sum_{j=1}^{m} \lambda_{ij} p(\omega_j|X) \qquad (1)$$

where

$X$ is an n-variate observation (feature) vector

$\{\omega_j, j=1, 2,..., m\}$ is the set of m classes

$\lambda_{ij}$ is the cost resulting from classifying into class i a pattern actually from class j

$p(\omega_j|X)$ is the conditional probability that, given observation X, its class is $\omega_j$

That is, $L_X(\omega_i)$ is the expected loss incurred if an observation X is classified as $\omega_i$. Commonly [2] $\lambda_{ij}$ is taken to be the "0-1 loss function," i.e.,

$\lambda_{ij} = 0, \quad i = j$ (no cost for correct classification)

$= 1, \quad i \neq j$ (unit cost for an error)

Then Eq. (1) becomes

$$L_X(\omega_i) = 1 - p(\omega_i|X) \qquad (2)$$

and an appropriate decision rule which will minimize $L_X(\omega_i)$ is:

Decide $X \in \omega_i$ if and only if

$$p(X|\omega_i)p(\omega_i) = \max_j p(X|\omega_j)p(\omega_j) \qquad (3)$$

where $p(X|\omega_i)$ is the probability density function for the obser-

vations associated with class $\omega_i$ and $p(\omega_i)$ is the a priori probability of class $\omega_i$. Thus the set of products $\{p(X|\omega_i)p(\omega_i), i=1, 2,..., m\}$ is a set of discriminant functions for the classification problem.

We now generalize this Bayes optimal approach to the case of a series of observations. It will be convenient to assume that observations are made at two times. Generalization to a larger number of observation times is straightforward.

Let $X_1 = X(t_1)$ and $X_2 = X(t_2)$ be n-variate random vectors, the pattern observations at times $t_1$ and $t_2$, respectively.

Let $\{\nu_i = \nu_i(t_1) \mid i=1,2,..., m_1\}$ be the set of possible classes at time $t_1$, and let $\{\omega_i = \omega_i(t_2) \mid i=1,2,..., m_2\}$ be the set of possible classes at time $t_2$.

We define a compound conditional average loss

$$L_{X_1X_2}(\omega_i) = \sum_{j=1}^{m_2} \lambda_{ij} p(\omega_j | X_1, X_2) \tag{4}$$

where $\lambda_{ij}$ is the cost resulting from classifying into class i, at time $t_2$, a pattern actually from class j. In this case $p(\omega_j | X_1, X_2)$ is the a posteriori probability that, given the observations $X_1$ at time $t_1$ and $X_2$ at time $t_2$, the class of the pattern at time $t_2$ is $\omega_j$.

Once again assuming a "0-1 loss function," Eq. (4) becomes

$$L_{X_1X_2}(\omega_i) = 1 - p(\omega_i | X_1, X_2) \tag{5}$$

which is minimized if we choose $\omega_i$ to maximize the a posteriori probability $p(\omega_i|X_1, X_2)$. Thus an appropriate set of discriminant functions for a Bayes optimal classification strategy is the set of a posteriori probabilities; i.e.

$$\left\{p(\omega_i|X_1, X_2), \quad i = 1,2,\ldots, m_2\right\}$$

As usual, however, we wish to derive a set of equivalent discriminant functions expressed in terms of class-conditional density functions and a priori probabilities as in Eq. (3). This may be accomplished proceeding as follows. First we write:

$$p(\omega|X_1, X_2) = \frac{p(\omega,X_1,X_2)}{p(X_1, X_2)} \tag{6}$$

For fixed $X_1$ and $X_2$, the denominator in Eq. (6) is constant. Let $c = 1/p(X_1, X_2)$ and write Eq. (6) as

$$\begin{aligned} p(\omega|X_1,X_2) &= cp(\omega,X_1,X_2) \\ &= c \sum_{\nu} p(X_1, X_2, \nu,\omega) \\ &= c \sum_{\nu} p(X_1, X_2|\nu,\omega)p(\nu,\omega) \\ &= c \sum_{\nu} p(X_1, X_2|\nu,\omega)p(\omega|\nu)p(\nu) \end{aligned} \tag{7}$$

The summation is over the classes which can occur at time $t_1$. The factor $p(X_1, X_2|\nu,\omega)$ is a joint class-conditional density; $p(\omega|\nu)$ may be interpreted as a transition probability (the probability that the class is $\omega$ at time $t_2$ given the class was $\nu$ at time $t_1$); and $p(\nu)$ is an a priori probability.

Thus, the multiobservational decision rule analogous to Eq. (3) is:

Decide $X_2 \in \omega_i$ if and only if

$$\sum_{k=1}^{m_1} p(X_1, X_2 | \nu_k, \omega_i) p(\omega_i | \nu_k) p(\nu_k) = \max_j \sum_{k=1}^{m_1} p(X_1, X_2 | \nu_k, \omega_j) p(\omega_j | \nu_k) p(\nu_k) \quad (8)$$

and the set of discriminant functions is the set of sums of products:

$$\left\{ \sum_{k=1}^{m_1} p(X_1, X_2 | \nu_k, \omega_i) p(\omega_i | \nu_k) p(\nu_k), \quad i=1,2,\ldots, m_2 \right\}. \quad (9)$$

## A "Cascade" Implementation

In practice, the terms in the discriminant functions must be estimated from "training samples." The most formidable job is estimating the $m_1 \cdot m_2$ joint class-conditional densities $p(X_1, X_2 | \nu_k, \omega_i)$, each of which is of dimension $2n$.[2] Clearly a large number of training samples will be required. When certain approximations can be justified, the situation is eased considerably. We shall now show that these approximations lead to a rather attractive model for a multitemporal classifier.

---

[2]The observation vectors need not be of the same dimensionality. If $X_1$ has $n_1$ components and $X_2$ has $n_2$ components, the $p(X_1, X_2 | \nu, \omega)$ is N-variate, where $N = n_1 + n_2$.

We are accustomed to assuming class-conditional independence in the spatial domain; i.e., given the class at a particular point, the random variable which is the measurement vector at that point is independent of the class or measurement vector at any other point. Applying this same idea to multitemporal measurements at a given point, we say that given the classes $\nu_k$ at $t_1$ and $\omega_i$ at $t_2$, the random variables $X_1$ and $X_2$ are independent. Then we can write

$$p(X_1, X_2 | \nu_k, \omega_i) = p(X_1 | \nu_k, \omega_i)\ p(X_2 | \nu_k, \omega_i) \tag{10}$$

and furthermore

$$\begin{aligned} p(X_1 | \nu_k, \omega_i) &\cong p(X_1 | \nu_k) \\ p(X_2 | \nu_k, \omega_i) &\cong p(X_2 | \omega_i) \end{aligned} \tag{11}$$

Imposing these conditions, it follows that

$$p(X_1, X_2 | \nu_k, \omega_i) = p(X_1 | \nu_k) p(X_2 | \omega_i).$$

The discriminant functions, Eq. (9), then become

$$\left\{ \sum_{k=1}^{m_1} p(X_1 | \nu_k) p(X_2 | \omega_i) p(\omega_i | \nu_k) p(\nu_k),\ i=1,2,\ldots, m_2 \right\} \tag{12}$$

From Eq. (12) we can model the discriminant function calculations as indicated in Figure 1, from which we derive the term "cascade classifier" to describe this multistage classifier.

## Simulation and Experimental Results

The cascade classifier model was programmed and applied to the analysis of a set of Landsat multispectral data. The data,

collected by the satellite on two successive passes, eighteen days apart, over Fayette County, Illinois (see Table 1), were geometrically registered at Purdue University's Laboratory for Applications of Remote Sensing. The objective of the analysis was to discriminate among the ground cover classes "corn", "soybeans", woods", and "other", where the last category was simply a catch-all consisting of water, pasture, fallow and other relatively minor ground covers. Each class was actually decomposed in the analysis process into a union of subclasses, each having a data distribution describable as approximately multivariate normal.[3]

To provide a baseline for comparison, the data from each of the passes was first analyzed separately. The a priori probabilities of the classes were approximated as being equal, and 557 test samples, independent of the training samples, were used to evaluate the results. As shown in Table 1(a) and (b), the performance of this conventional maximum likelihood classifier was 68% correct for the June 29, 1973 data, and 72% correct for the July 17, 1973 data.

To implement the cascade analysis, it was assumed unlikely that the ground cover would change identity over so short a time span. Accordingly, the transition probabilities were estimated as follows:

$$p(\omega_i|\nu_k) = 0.8 \quad \text{for } \omega_i = \nu_k, \tag{13a}$$

and all other transition probabilities were set equal and such that

---

[3]All probability densities were assumed to be multivariate normal (Gaussian), characterized by mean vector and covariance matrix.

$$\sum_{\substack{i \\ \omega_i \neq \nu_k}} p(\omega_i | \nu_k) = 0.2. \qquad (13b)$$

Again the a priori probabilities were assumed equal and the same test samples were used to evaluate the results.

The results of this multitemporal classification, Table 1(c), were substantially better than either of the unitemporal analyses. The overall results were 84% correct. In addition, the performance for each class was better than the best attained for the class in either of the unitemporal analyses. The unitemporal and multitemporal results are compared in Figure 2.

The results can be sensitive, however, to the specification of the transition probabilities and a priori probabilities. This is demonstrated in the following experiment.

Landsat data from two passes over Grant County, Kansas, were analyzed in a manner similar to that used for the Fayette County data. In this case, the two passes were separated by more than two months and a different set of classes was involved (Table 2). The transition probabilities were specified as in Eq. (13a) and (13b); equal a priori probabilities were assumed.

As shown in Table 2 and Figure 3, in this case the overall performance of the multitemporal cascade classifier was only marginally better than the best unitemporal result. A closer look at the class-by-class results is revealing. The largest detractors from the multitemporal results were the classes "alfalfa" and "pasture." In both of these cases, the unitemporal results for the second pass were substantially lower than those obtained in the first pass. (There are physical explanations for why this is reasonable, but this is not germane to our exploration of classifier behavior.)

Let us examine the impact that the relatively arbitrary assignment of transition probabilities has on the classification results. In case the actual transition probabilities are not known (which was true for the cited examples), the assignment can be made anywhere between two extremes. On the one hand, it could be assumed that

$$p(\omega_i|\nu_k) = \frac{1}{m_1} \quad , \quad k = 1,2,\ldots, m_1$$

i.e., equiprobable transitions. Then the discriminant functions have the form

$$\sum_{k=1}^{m_1} p(X_1|\nu_k)p(X_2|\omega_i)\frac{1}{m_1} \; p(\nu_k)$$

$$= \frac{1}{m_1} \; p(X_2|\omega_i) \sum_{k=1}^{m_1} p(X_1|\nu_k)p(\nu_k)$$

$$= \frac{1}{m_1} \; p(X_2|\omega_i)p(X_1).$$

Since $\frac{1}{m_1}$ and $p(X_1)$ will be common to each of the discriminant functions, the decision will depend only on $p(X_2|\omega_i)$ and will be independent of the first-stage results.

On the other hand we could make $p(\omega_i|\nu_i) = 1$ and $p(\omega_i|\nu_j) = 0$, $j \neq i$. Then the discriminant functions become

$$p(X_1|\nu_i)p(X_2|\omega_i)p(\nu_i).$$

Thus, in a sense, the contributions from the two stages are weighted equally.

There is no way to make the first stage input dominate the second stage.

In view of these considerations, another classification of the Grant County data was performed. In this case, the transition probabilities $p(\omega_i|v_i)$ were set equal to unity for the "alfalfa" and "pasture" classes in order to give as much strength as possible to the first stage results. Table 3 and Figure 3 show the outcome of this classification. The confusing influence resulting from the second stage data has been reduced.

It is interesting to compare the results obtained using the cascade classifier to results produced by a "conventional" maximum likelihood classifier using all of the multitemporal features simultaneously. To perform the latter classifications, equal a priori probabilities were assumed. The results were:

Fayette County: 80.8 percent correct

Grant County: 64.1 percent correct

It is curious that neither of these results is any better than the cascade classifier results achieved. It is possible that these slightly poorer results represent the price paid for having to estimate 8-dimensional statistics as opposed to 4-dimensional statistics in the face of limited training data.

## Discussion and Conclusions

The approach we have adopted for classifying data in a non-stationary environment was based on application of classical statistical decision theory in a straightforward manner. However, we used the conditions of the problem to approximate some of the statistical quantities involved. This step simplified the interdependencies of the data involved and led to a "cascade classifier"

model. In the time-varying environment, this model is seen to:

(1) Successfully incorporate the temporal information in the classification process, resulting in improved classification accuracy;

(2) Reduce the dimensionality of the probability functions used and thereby make less stringent demands with respect to the size of the training set required;

(3) Facilitate distribution of the computational load over time.

Each time a set of observations becomes available, discriminant functions are calculated which can be used, if desired, to make a classification. However, the values of the discriminant functions are also passed along and contribute to a new set of discriminant functions calculated when the next set of observations is obtained. Although we have demonstrated the use of the cascade model only for the case of two stages, extension to an arbitrary number of stages presents no difficulty.

The prospective user of this approach should be aware that a casual implementation of the likelihood computers may result in computational difficulties of two sorts: loss of precision and very large computation times as compared with, say, a conventional Gaussian maximum likelihood classifier. Both of these difficulties can be overcome or at least substantially reduced by appropriate measures (scaling, ignoring zero terms, etc.) in carrying out the likelihood computations.

## Acknowledgements

The author wishes to thank Mr. Carlos A. Pomalaza for programming the cascade classifier model and testing it with the remote sensing data. This research was supported in part by NASA Contract NAS9-14970.

## References

[1] Landgrebe, D.A., "The Quantitative Approach: Concept and Rationale," Chapter 1 in P.H. Swain and S.M. Davis, eds., Remote Sensing: The Quantitative Approach, McGraw-Hill International Book Co., Inc., 1978.

[2] Nilsson, N.J., Learning Machines, McGraw-Hill Book Co., Inc., 1965.

Table 1. Test results for classification of the Fayette County, Illinois, data.

(a) June 29, 1973 data

| Group | No. of Samples | Percent Correct | No. of Samples Classified into CORN | OTHERS | SOYBEAN | WOODS |
|---|---|---|---|---|---|---|
| CORN | 186 | 65.1 | 121 | 36 | 24 | 5 |
| OTHERS | 100 | 40.0 | 33 | 40 | 22 | 5 |
| SOYBEAN | 227 | 82.4 | 10 | 30 | 187 | 0 |
| WOODS | 44 | 72.7 | 0 | 4 | 8 | 32 |
| TOTAL | 557 | | 164 | 110 | 241 | 42 |

OVERALL PERFORMANCE = 68.2 percent correct

(b) July 17, 1973

| Group | No. of Samples | Percent Correct | No. of Samples Classified Into CORN | OTHERS | SOYBEAN | WOODS |
|---|---|---|---|---|---|---|
| CORN | 186 | 89.2 | 166 | 16 | 1 | 3 |
| OTHERS | 100 | 45.0 | 38 | 45 | 15 | 2 |
| SOYBEAN | 227 | 73.6 | 24 | 36 | 167 | 0 |
| WOODS | 44 | 56.8 | 4 | 9 | 6 | 25 |
| TOTAL | 557 | | 232 | 106 | 189 | 30 |

OVERALL PERFORMANCE = 72.4 percent correct

(c) Multitemporal results (cascade classifier)

| Group | No. of Samples | Percent Correct | No. of Samples Classified Into CORN | OTHER | SOYBEAN | WOODS |
|---|---|---|---|---|---|---|
| CORN | 186 | 90.3 | 168 | 11 | 4 | 3 |
| OTHERS | 100 | 48.0 | 29 | 48 | 20 | 3 |
| SOYBEAN | 227 | 94.3 | 3 | 10 | 214 | 0 |
| WOODS | 44 | 84.1 | 0 | 5 | 2 | 37 |
| TOTAL | 557 | | 200 | 74 | 240 | 43 |

OVERALL PERFORMANCE = 83.8 percent correct

Table 2. Test results for classification of the Grant County, Kansas, data.

(a) May 9, 1974

| Group | No. of Samples | Percent Correct | No. of Samples Classified Into ALFALFA | CORN | FALLOW | PASTURE | WHEAT |
|---|---|---|---|---|---|---|---|
| ALFALFA | 58 | 84.5 | 49 | 0 | 0 | 0 | 9 |
| CORN | 428 | 57.0 | 0 | 244 | 183 | 1 | 0 |
| FALLOW | 526 | 54.4 | 0 | 196 | 286 | 36 | 8 |
| PASTURE | 1513 | 52.6 | 127 | 148 | 220 | 796 | 227 |
| WHEAT | 913 | 82.5 | 97 | 17 | 0 | 49 | 767 |
| TOTAL | 3455 | | 273 | 605 | 689 | 882 | 1006 |

Overall Performance = 62.0 percent correct

(b) July 20, 1974

| Group | No. of Samples | Percent Correct | No. of Samples Classified Into ALFALFA | CORN | FALLOW | PASTURE | WHEAT |
|---|---|---|---|---|---|---|---|
| ALFALFA | 58 | 5.2 | 3 | 3 | 0 | 10 | 42 |
| CORN | 428 | 53.0 | 15 | 227 | 105 | 15 | 66 |
| FALLOW | 526 | 62.9 | 0 | 113 | 331 | 5 | 77 |
| PASTURE | 1513 | 42.4 | 64 | 329 | 213 | 641 | 266 |
| WHEAT | 913 | 76.2 | 22 | 108 | 33 | 58 | 709 |
| TOTAL | 3455 | | 104 | 780 | 682 | 729 | 1160 |

Overall Performance = 55.3 percent correct

(c) Multitemporal results (cascade classifier)

| Group | No. of Samples | Percent Correct | Number of samples classified Into ALFALFA | CORN | FALLOW | PASTURE | WHEAT |
|---|---|---|---|---|---|---|---|
| ALFALFA | 58 | 41.4 | 24 | 0 | 0 | 2 | 32 |
| CORN | 428 | 59.6 | 5 | 255 | 165 | 1 | 2 |
| FALLOW | 526 | 76.4 | 0 | 107 | 402 | 2 | 15 |
| PASTURE | 1513 | 46.3 | 101 | 205 | 224 | 701 | 282 |
| WHEAT | 930 | 88.3 | 77 | 19 | 0 | 13 | 821 |
| TOTAL | 3455 | | 207 | 586 | 791 | 719 | 1152 |

Table 3. Cascade classifier results for adjusted transition probabilities (Grant County data).

| Group | No. of Samples | Percent Correct | Number of samples classified Into ALFALFA | CORN | FALLOW | PASTURE | WHEAT |
|---|---|---|---|---|---|---|---|
| ALFALFA | 58 | 94.8 | 55 | 0 | 0 | 0 | 3 |
| CORN | 428 | 70.3 | 5 | 301 | 122 | 0 | 0 |
| FALLOW | 526 | 68.1 | 0 | 139 | 358 | 7 | 22 |
| PASTURE | 1513 | 48.1 | 105 | 211 | 195 | 727 | 275 |
| WHEAT | 930 | 89.1 | 82 | 9 | 0 | 10 | 829 |
| TOTAL | 3455 | | 247 | 660 | 675 | 744 | 1129 |

Overall Performance = 65.7 percent correct

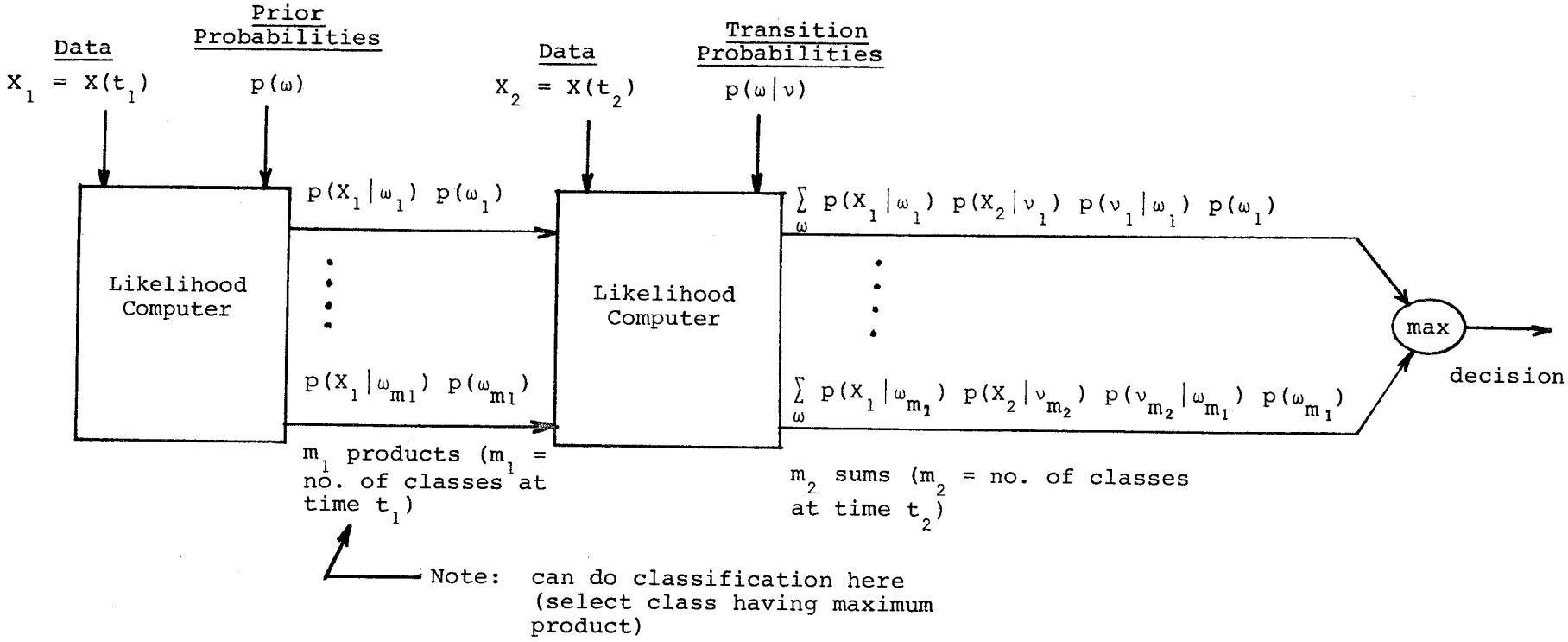


Figure 1. The cascade classifier model.

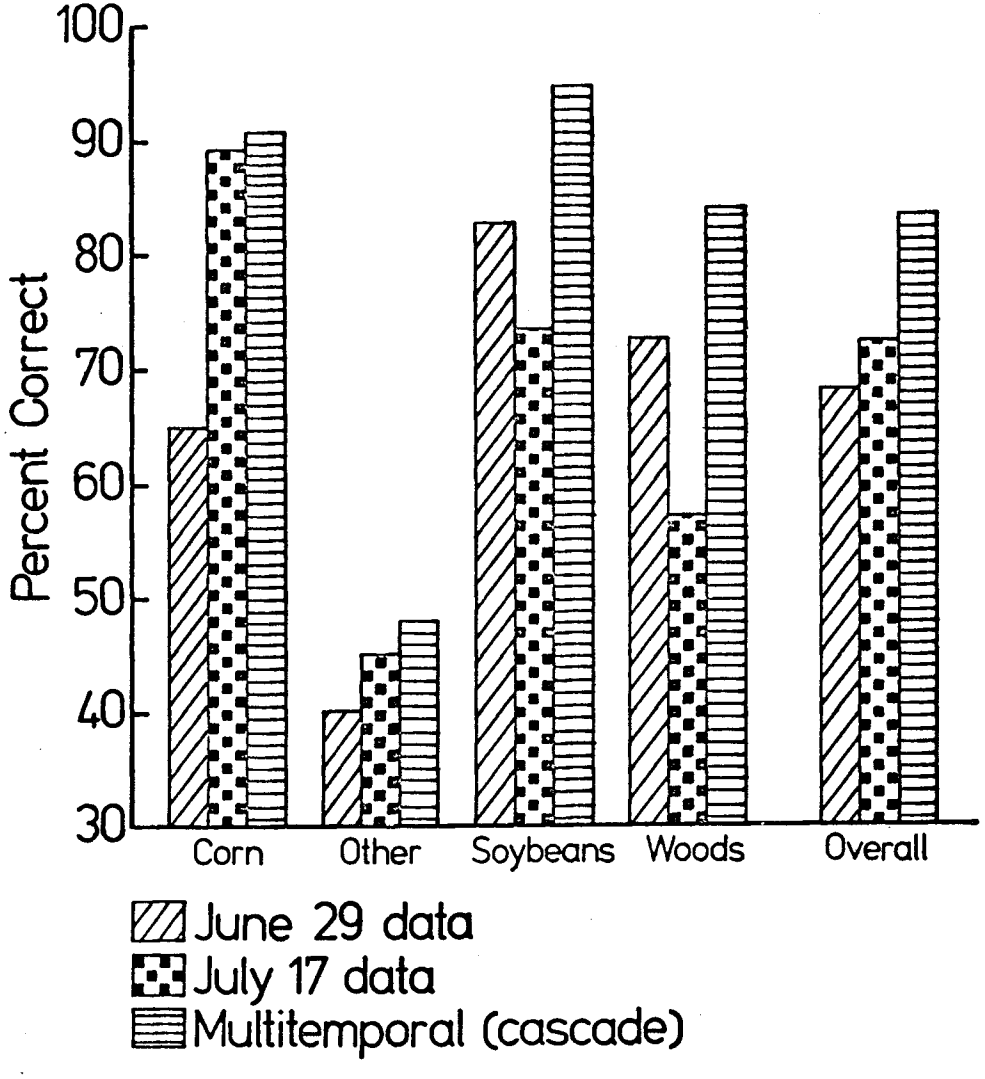


Figure 2. Test results for Fayette County data.

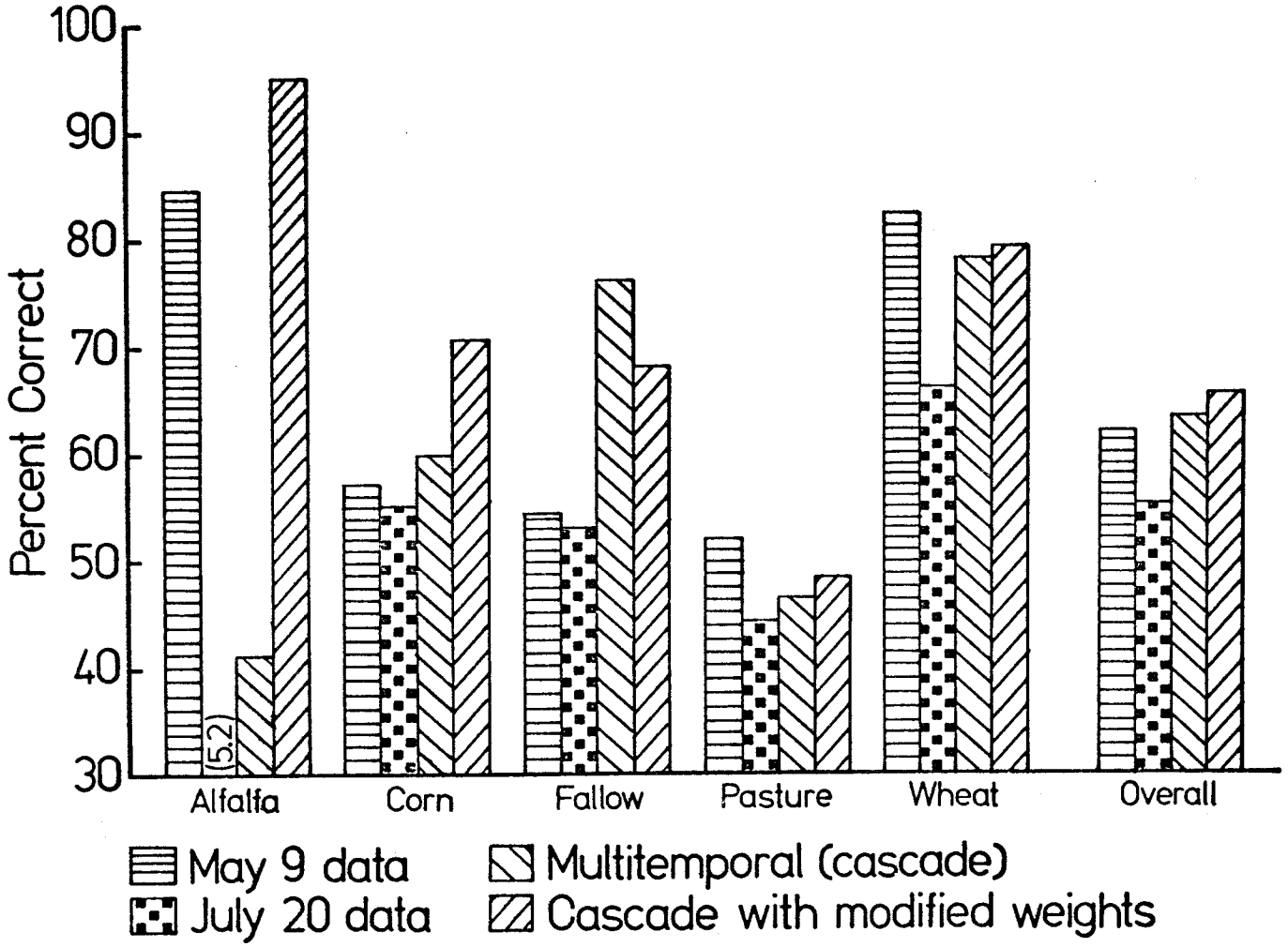


Figure 3. Test results for Grant County data.