JOURNAL OF THE TENNESSEE ACADEMY OF SCIENCE
Volume 46, Number 1, January, 1971

# A NOTE ON SOME PRINCIPLES OF INFORMATION THEORY AND SIGNAL DETECTION APPLIED TO MEASUREMENT SYSTEMS

HORACE E. WILLIAMS
*Vanderbilt University*
*Nashville, Tennessee* 37203

ABSTRACT

Some concepts from information theory employed by engineers working in the field of signal detection provide useful objective measures of risk and information content for analyzing problems where an investigator would like to infer some one of a discrete set of events or states of nature using a discrete set of responses or signal outputs as a basis for the inference. There appear to be fruitful areas of application of these concepts to inference problems encountered in the social and industrial sciences. In particular, the "entropy" and "information" functions provide a valuable enrichment to a straightforward probabilistic analysis of such problems.

In recent years Information Theory and Signal Detection, once the exclusive domain of communications and systems engineers, has been stirring interest in other areas of engineering and the sciences. Some of the ideas from these fields appear adaptable to the social and industrial sciences and have caused considerable discussion among persons in these fields.

Information theory was principally originated by Claude Shannon in 1948 and 1949 and is a branch of probability theory. It also borrows heavily from concepts in thermodynamics. The power of Information Theory comes from its ability to quantify the predictability of a discrete system and to provide a device for measuring the amount of information derivable from observing a given system and for discriminating among possible alternatives in the system.

The purpose of this discussion is to briefly review the basic concepts of Information Theory and Signal Detection in a discrete system environment and to present some feasible areas in which these ideas may materially assist in the analysis of a problem. One should not expect these ideas to open new vistas of research, rather the hope is that it will provide means for improving the clarity and preciseness of an analysis of a given problem.

## THE GENERAL PROBLEM

Given a discrete set of possible events or states of nature $E_1, E_2, E_3, \ldots, E_n$, an investigator desires to infer some one of the events by using previous knowledge and obtainable data. Associated with the events are a discrete set of related signal outputs $V_1, V_2, \ldots, V_m$ that may be measured (perhaps system responses or activities might be a meaningful alternative designation than signal outputs). The tasks are to determine an optimum decision process for the given system for inferring events when given the signal outputs, and to establish criteria for comparing feasible decision schemes and available signal sets.

If one may assume from theoretical considerations or establish from previous experience a prior probability distribution for the events, $P(E_i)$, and if further one may establish the conditional probabilities of the signal outputs $V_j$ given each event $E_i$, $P(V_j \mid E_i)$, the first part of the analysis may be carried out as a straightforward problem in conditional probabilities. Using Bayes Theorem, the back probabilities, i.e. the conditional probabilities of the events $E_i$ given each signal output $V_j$, $P(E_i \mid V_j)$, may be computed. These back probabilities may then be used for inferring the likelihoods of the given events given various detector outputs. It may be shown that the minimum error decision rule is to infer that event for each detector output $V_j$ for which the conditional probability $P(E_i \mid V_j)$ is maximum. [Abramson, 1963; Middleton, 1960]. This rule of course corresponds nicely to one's intuitive evaluation of such a procedure.





This analytic procedure may be enriched by ideas from information theory. In particular, the "entropy" and the "average mutual information" are important concepts in the description of such a system. These provide objective measures of the average risk inherent in the inference system and of the average information provided by the system. They also allow one to make objective comparisions of several possible competing signal sets or to assess the efficiency of using combinations of several signal sets.

The event entropy of a discrete system may be defined as

$$H(E) = \sum_{i=1}^{n} P(E_i) \log_2 \frac{1}{P(E_i)} \quad (1)$$ *

This function assesses the "no measurement risk" of the system, that is the risk associated with making an inference on the basis of prior probabilities alone without benefit of any measurements. It is a sensible and useful measure in that it has the following very desirable properties:

a) $H(E) = 0$ if one of the prior probabilities is 1 and the remainder are zero. This property is certainly intuitively pleasing since one would be disposed to think of such a case as being a no risk inference. This constitutes a minimum value for $H(E)$.

b) $H(E)$ achieves its maximum value and is equal to $\log_2 n$ when all the prior probabilities are equal. Again this agrees favorably with intuition since it represents the case of pure guessing which one would like to think represents maximum risk.

The "one measurement risk" or "entropy" is defined as

$$H(E|V) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(E_i, V_j) \log_2 \frac{1}{P(E_i \mid V_j)} \quad (2)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} P(E_i)\, P(V_j \mid E_i) \log_2 \frac{1}{P(E_i \mid V_j)}.$$ **

This function provides a reasonable assessment of the average risk associated with making an inference which incorporates the results of measuring the output of the signal set $V_j$. Since the term risk is perhaps more meaningful in this context than entropy, we shall henceforth use that term. The inequalities $0 \leq H(E|V) \leq H(E)$ provide bounds for the function.

The "average mutual information" function is defined as (3) $I(E;V) = H(E) - H(E|V)$. This function provides a measure of the average information about the events $E_i$ provided by the signal set $V_j$, or stated alternatively, the average mutual information associated with the process of inferring E from the output of V. The information function is found to posess some useful and desirable properties:

a) $I(E;V)$ is maximum in the case of an ideal signal set (a set which permits one to discriminate perfectly between the events $E_i$). For the ideal signal set $H(E|V) = 0$ so that max $I(E;V) = H(E)$. Such a system contains the maximum possible information and reduces the risk of inference to zero.

b) $I(E;V) = 0$ in a case in which the signal set so poorly discriminates between the events that all inferences are equilikely. In such a system the measurements of $V_j$ have provided no usable information about the events $E_i$. The minimum value for $I(E;V)$ is zero, establishing the bounds $0 \leq I(E;V) \leq H(E)$.

* Logarithms to bases other than 2 may be used but interpretations and tables of values for base 2 have been extensively developed. Abramson, 1963; Middleton, 1960.

** The notation $P(E_i, V_j)$ refers to the joint probability of event $E_i$ and signal output $V_j$.

The function I(E;V) may thus be regarded as a measure of correlation or relationship between the events $E_i$ and the responses $V_j$. The quantity 1.3863 k I(E;V), where k is the sample size used to estimate the prior probabilities of the events E, is in effect equivalent to the value of chi-square that one would use to test the hypothesis that the sets E and V are independent (or stated in other words, the hypothesis that the system provides a statistically significant amount of information). I(E;V) may also be interpreted as a measure of the discriminatng ability of the response set V about the events E.

Let us now examine some examples of the possible use of the principles discussed. These examples will be limited to quite simple cases in the interests of conciseness of presentation; however, the ideas illustrated may be easily extended to more complex situations.

EXAMPLES.

A) Suppose that the superintendent of a factory is considering the purchase of a piece of testing equipment to assist in the classification of defects in the manufactured product. Further suppose that many of the defects are correctible. From past statistical records it is estimated that 60% of the defects are correctible while 40% are not. The piece of equipment is then subjected to controlled field testing to estimate its capabilities using laboratory prepared samples. On the basis of this suppose that of samples which contained correctible defects the equipment indicated 58% of them as being correctible and 42% of them as being non-correctible. Of the non-correctible samples the equipment indicated 16% of them as being correctible and 84% as being non-correctible. Superficially it appears that the test equipment may not be of any specific benefit.

The superintendent feels that without the purchase of the test equipment he will have an efficiency ratio of 60% if he attempts to have all defects corrected. The question he wishes to answer is whether he may economically increase his efficiency if he employs the test equipment.

One may represent this problem by a tree diagram as shown in Figure 1.

$E_1$ - product has a correctible defect.

$E_2$ - product has an uncorrectible defect.

$V_1$ - test equipment indicates a correctible defect.

$V_2$ - test equipment indicates an uncorrectible defect.

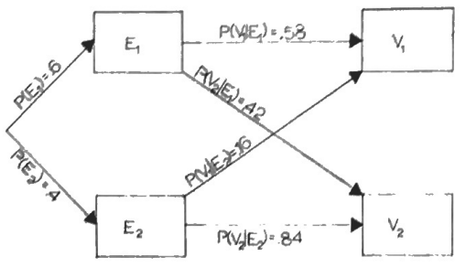



FIGURE 1

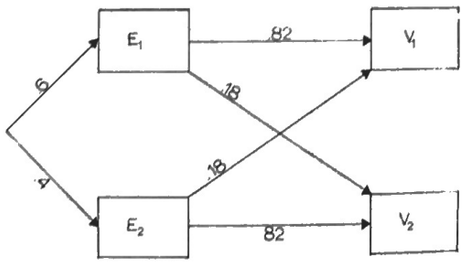



FIGURE 2



The back probabilities may now be computed using Bayes Theorem. Bayes Theorem states:

$$P(E_i \mid V_j) = \frac{P(E_i)\ P(V_j \mid E_i)}{\sum_{i=1}^{n} P(E_i)\ P(V_j \mid E_i)} = \frac{P(E_i, V_j)}{P(V_j)} \tag{4}$$

Applying equation (4) to the diagram of Figure 1, we obtain:

$$P(E_1|V_1) = \frac{(.6)\ (.58)}{(.6)\ (.58) + (.4)\ (.16)} = \frac{.348}{.412} = .845,$$

$$P(E_2 \mid V_1) = 1 - P(E_1 \mid V_1) = .155,$$

$$P(E_1 \mid V_2) = \frac{(.6)\ (.42)}{(.6)\ (.42) + (.4)\ (.84)} = \frac{.252}{.588} = .429,$$

$$P(E_2 \mid V_2) = 1 - P(E_1 \mid V_2) = .571.$$

It is thus seen by using the test equipment the probability of identifying a correctible defect in the product is increased from .60 to .845, and the probability of identifying an uncorrectible defect is increased from .40 to .571. The savings resulting from the increased efficiency may then be compared with the depreciated purchase cost of the equipment. The no measurement risk is $H(E) = .6 \log 1/.6 + .4 \log 1/.4 = .972$. This is a measure of the risk taken when not using the testing equipment and attempting to correct all defects. The one measurement risk is $H(E|V) = (.6)\ (.58) \log (1/.845) + (.6)\ (.42) \log (1/.429) + (.4)\ (.16) \log (1/.155) + (.4)\ (.84) \log (1/.571) = .838$. This gives a measure of the risk taken when using the testing equipment. The information is $I(E;V) = .972 - .838 = .134$.

If the superintendent has available another, perhaps better, piece of testing equipment from a competing firm that he may purchase, he may compare the information content of the two. Suppose the tree for the second piece of equipment is represented by the diagram in Figure 2.

The back probabilities are:

$P(E_1|V_1) = .875$, $P(E_1|V_2) = .244$,

$P(E_2|V_1) = .125$, $P(E_2|V_2) = .756$,

$H(E) = .972$, $H(E|V) = .661$,

$I(E;V) = .311$.

The second piece of equipment is shown to have a demonstrably better information content than the first and both are less risky than employing no testing. Of course a sensible conclusion for the superintendent, if time and expense do not interfere (frequently in complex situations this is cruicial), might be to investigate using both. If he did the risk would be .580 and the information .392. The method for analysing such multiple signal sets will be discussed later.

Although the example is simplified for a case in which there are only two states of nature, this type of analysis may easily be extended to a situation in which there are several states of nature. For instance, there might be several categories of defects that need to be identified and a piece of testing equipment used which attempts to give responses which will discriminate between the several types.

B) As a second example, suppose a political scientist is interested in inferring attitudes of individuals in a community regarding a socio-political issue. He defines, let us say, three attitude categories: favorable, neutral, and unfavorable. He is further interested in determining if a knowledge of available background data such as sex, race, political party affiliation, employment status and others influence the inferences and if so which exert the most influence. He may survey a sample of individuals in the community and use the sample data as a basis for the establishment of probabilities for use in an information theory model. He may then set up tree diagrams for the possible relationships as illustrated in Figure 3 - depicting attitude vs. employment status.

The average mutual information may then be evaluated for each relationship, tested for significance, and this used as a basis for determining the relative discriminating power of a given piece of background data.

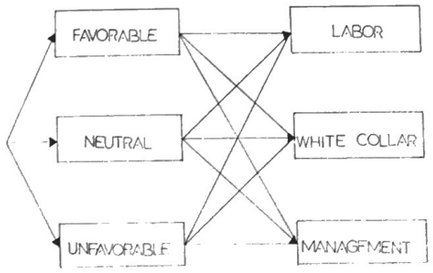



FIGURE 3

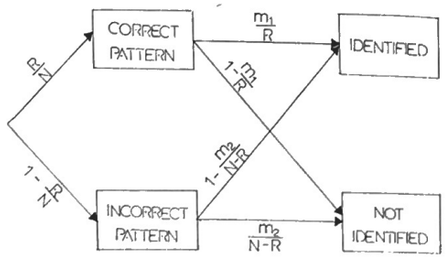



FIGURE 4

C) As a third application, it appears feasible to employ the information theory concepts in certain situations involving the training of persons for use as detectors. Suppose, for instance, that in a factory persons are being trained as inspectors on the production line and are to identify a certain type of pattern within a short time interval. A system for evaluating the training program and for comparing the relative improvements of individuals to whom the training has been administered is to be designed. An efficiency test of the following type could be employed. An individual is randomly exposed to N patterns of which a predetermined number R are correct and the remainder incorrect. He is asked to make an identification for each pattern. Suppose the individual identified properly $m_1$ of the correct patterns and $m_2$ of the incorrect patterns. The diagram shown in Figure 4 represents his performance. The average mutual information may be computed for this test using the principles developed and be used as a measure of efficiency. A test may be administered to a group of trainees prior to training and a parallel test administered subsequent to training, and the mean information gain used as a measure of training efficiency. The pre-test vs. post-test information gain for an individual may be used as a measure of relative improvement for the individual.

EXTENSIONS.

Let us return now to the question raised earlier about the use of two or more independent signal sets in conjunction for making inferences. To illustrate concretely the analytical technique employed, consider the signal systems shown in Figures 1 and 2 and assume that the test equipment represented by Figure 1 is used first and then the equipment represented by Figure 2 is used. Denote the responses to these by $V_1$, $V_2$ and $Q_1$, $Q_2$ respectively (interchanging the order of the two signal sets is immaterial to the results and the ordering may simply be chosen at the convenience of the investigator). The probabilities of responses $V_1$ and $V_2$ may now be computed by tracing all possible paths in Figure 1 leading to these outcomes:



$$P(V_1) = (.6)\ (.58) + (.4)\ (.16) = .412,$$

$$P(V_2) = (.6)\ (.42) + (.4)\ (.84) = .588.$$

Next we construct new tree diagrams for each of the possible outcomes of test 1 using the back probabilities $P(E_1|V_1)$, $P(E_2|V_1)$, $P(E_1|V_2)$, $P(E_2|V_2)$, previously computed.

Assuming response $V_1$ for test 1 we obtain the tree shown in Figure 5. The back probabilities, the risk, and the information may be computed for this tree in the same fashion as has been done previously:

$P(E_1|V_1,Q_1) = .96,$ $P(E_2|V_1,Q_1) = .04,$

$P(E_1|V_1,Q_2) = .545,$ $P(E_2|V_1,Q_2) = .455,$

$H(E|V_1) = .622,$ $H(E|V_1,Q) = .488,$

$I(E;Q|V_1) = .174$ (Partial information of test 2 given response $V_1$ on test 1).

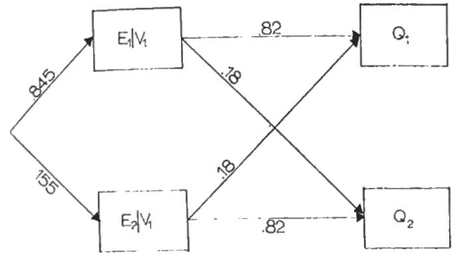



FIGURE 5

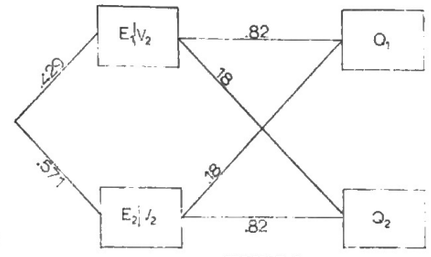



FIGURE 6

Assuming response $V_2$ for test 1, we obtain the tree of Figure 6. The statistics for this case are:

$P(E_1|V_2,Q_1) = .774,$ $P(E_2|V_2,Q_1) = .226,$

$P(E_1|V_2,Q_2) = .140,$ $P(E_2|V_2,Q_2) = .860,$

$H(E|V_2) = .985,$ $H(E|V_2Q) = .672,$

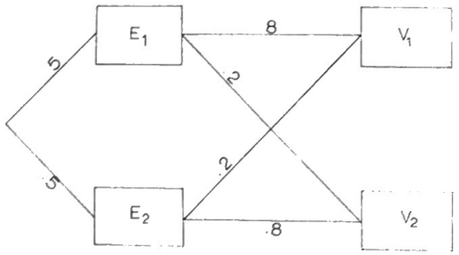
$I(E;Q|V_2) = .313$ (Partial information of test 2 given response $V_2$ on test 1).

The average risk for the two is evaluated in equation (5) by computing the weighted mean of the risks from Figures 5 and 6 using $P(V_1)$ and $P(V_2)$ as the weights. The mutual information, $I(E;V,Q)$, is then found as shown in equation (6).

(5) $$H(E|V,Q) = (.412)\ (.448) + (.588)\ (.672) = .580.$$

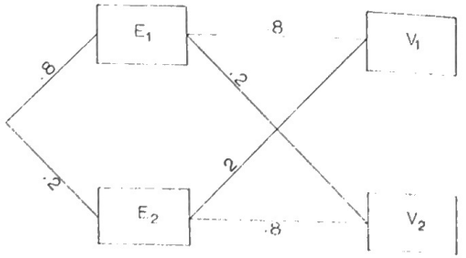
(6) $$I(E;V,Q) = H(E) - H(E|V,Q) = .972 - .580 = .392.$$

Another extension to the procedure may be found in certain more uniform situations by applying sequential analysis techniques. Suppose that one has available a relatively large reservoir of independent tests that are reasonably easy and cheap to apply, are approximately the same discriminating ability, and the prior probabilities may be assumed equal, $P(E_1) = P(E_2) = .50$. It is desired to administer the tests sequentially until an inference about either $E_1$ or $E_2$ is reached at a predetermined significance level, say .95. The procedure will be illustrated by reference to Figures 7, 8 and 9.




$P(E_1 | V_1) = .8$
$P(E_2 | V_1) = .2$
$P(E_1 | V_2) = .2$
$P(E_2 | V_2) = .8$

FIGURE 7




$P(E_1 | V_1) = .94$
$P(E_2 | V_1) = .06$
$P(E_1 | V_2) = .5$
$P(E_2 | V_2) = .5$

FIGURE 8

The first test is represented by Figure 7. If the response is $V_1$ then the second test is represented by Figure 8. If the response is again $V_1$ then the third test is represented by Figure 9. If once again the response is $V_1$ then we may infer $E_1$ with a .985 probability. Thus the response sequence $[V_1, V_1, V_1]$ establishes a decision within the predetermined significance level. If on the other hand the response to test 1 is $V_2$ then the second test is represented by Figure 8 with the indices reversed. If now the second response is $V_2$ then the third test is represented by Figure 9 with the indices reversed. If the response to the third test is again $V_2$ we may infer $E_2$ with a .985 probability, and the response sequence $[V_2, V_2, V_2]$ leads to an acceptable decision. If the response to the first test is $V_1$ and the response to the second test is $V_2$ then the third test is again represented by Figure 7 and from there we see that the sequence $[V_1, V_2, V_1, V_1, V_1]$ allows us to infer $E_1$. By tracing various response sequences through the given diagrams one may establish a catalogue of sequences leading to decisions at the given significance level. For instance we find:

$[V_2, V_1, V_1, V_1, V_1]$ decide $E_1$,

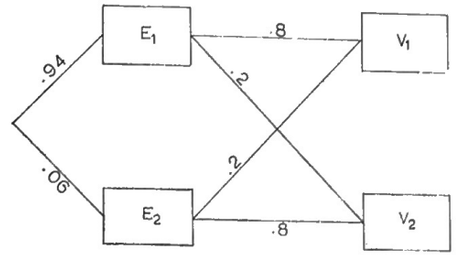
$[V_2, V_2, V_1, V_2, V_2]$ decide $E_2$,

$[V_1, V_1, V_2, V_1, V_2, V_1, V_1]$ decide $E_1$,

$[V_2, V_1, V_2, V_2, V_2]$ decide $E_2$.



An obvious difficulty inherent to such a sequential scheme of inference is that there is a small but finite possibility that the response sequence generated will never converge to a decision at the appropriate significance level. So one needs to establish a maximum number of tests that may be feasibly measured within limitations of time and expense and if this maximum is reached choose the best available decision. If for instance our maximum was seven tests in the current example and the sequence encountered was $[V_1, V_1, V_2, V_2, V_1, V_1, V_2]$ the best available decision is to infer $E_1$ with a probability .80.




$P(E_1 | V_1) = .985$
$P(E_2 | V_1) = .015$
$P(E_1 | V_2) = .8$
$P(E_2 | V_2) = .2$

FIGURE 9

## ADDITIONAL CONCEPTS OF INFORMATION THEORY.

The total information from all sources of a system such as those described may be computed by $H(E,V) = H(E) + H(V) - I(E;V)$. The "equivocation" of the system may subsequently be computed as

(7) $$H_V(E) = H(E,V) - H(V)$$

This gives a measure of the average amount of information about the events E that is unrelated or unpredictable from the responses V. The "noise" of the system is

(8) $$H_E(V) = H(E,V) - H(E)$$

The "noise" may be interpreted as the average amount of information resulting from the responses V that is of no value in discriminating between the events E.

The ratio of the information to the noise, $\frac{I(E;V)}{H_E(V)}$, is often referred to as the "signal to noise" ratio and is a useful comparative measure of efficiency or discriminating ability of a discrete response system in that a good system should have a relatively high information content and a relatively low noise level.

A characteristic of information systems is that usually there is an upper limit to the amount of average mutual information I(E;V) called by communication engineers "the channel capacity" C. C measures the ultimate predictability or discriminating ability of the system. C may in certain regular situations be calculated analytically; in other situations, however, it is only practical to estimate C by numerical methods. [Reza, 1961] .

The error probability of a Bayesian decision rule as outlined in this discussion may be computed by

$$P_b = 1 - \sum_j \max_i \left[P(V_j \mid E_i) \quad P(E_i)\right] . \tag{9}$$

Using error probabilities one may establish upper and lower bounds for the information I(E;V) for such a system. [Chu and Chueh, 1966] . These concepts may provide useful results in certain applications where one would like to establish a feasible range for the discriminating power of an information system.

## References


Abramson, Normal. *Information Theory and Coding.* New York: McGraw-Hill Inc., 1963.

Buckley, Walter (Editor). *Modern Systems Research for the Behavioral Scientist.* Chicago: Aldine Publishing Company, 1968.

Chu, J. T. and Chueh, J. C. Inequalities Between Information Measures and Error Probability. *Journal of the Franklin Institute,* 1966, 282, 121-125.

Churchman, C. West. *Prediction and Optimal Decision.* Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1961.

Kullback, Solomon. *Information Theory and Statistics.* New York: John Wiley and Sons, 1959.

Middleton, David. *An Introduction to Statistical Communications Theory.* New York: McGraw-Hill Inc., 1960.

Reza, F. M. *An Introduction to Information Theory.* New York: McGraw-Hill Inc., 1961.

Shannon, C. E. and Weaver, W. *The Mathematical Theory of Communication.* Urbana, Illinois: Univ. Illinois Press, 1949.