Methods For Combining Experts' Probability Assessments

Robert A. Jacobs
Department of Brain and Cognitive Sciences, University of Rochester,
Rochester, NY 14627 USA

This article reviews statistical techniques for combining multiple probability distributions. The framework is that of a decision maker who consults several experts regarding some events. The experts express their opinions in the form of probability distributions. The decision maker must aggregate the experts' distributions into a single distribution that can be used for decision making. Two classes of aggregation methods are reviewed. When using a supra Bayesian procedure, the decision maker treats the expert opinions as data that may be combined with its own prior distribution via Bayes' rule. When using a linear opinion pool, the decision maker forms a linear combination of the expert opinions. The major feature that makes the aggregation of expert opinions difficult is the high correlation or dependence that typically occurs among these opinions. A theme of this paper is the need for training procedures that result in experts with relatively independent opinions or for aggregation methods that implicitly or explicitly model the dependence among the experts. Analyses are presented that show that $m$ dependent experts are worth the same as $k$ independent experts where $k \leq m$. In some cases, an exact value for $k$ can be given; in other cases, lower and upper bounds can be placed on $k$.

1 Introduction

Biological perceptual systems are often conceptualized as containing two stages of processing: An early stage detects individual stimulus features; a later stage combines the detected features into abstract representations that facilitate thought, decision making, or action. Graham (1989, p. vii), for example, writes, "Visual perception can be described crudely as a two-part process: First, the visual system breaks the information in the visual stimulus into parts, and then second, the visual system puts the information back together again (not in its original form, however)." Graham goes on to ask, "Why should there be this analysis followed by a synthesis?" and observes that "current visual science knows much more about the parts than about the subsequent computations that put them back together again."
In recent years there has been an increase in the number of investigations into the processes responsible for combining the outputs of low-level feature analyzers. These processes are being studied from the perspectives of neuroscience (e.g., Stein and Meredith 1993; Zeki 1993), psychophysics (e.g., Nakayama and Shimojo 1990; Trueswell and Hayhoe 1993), and computational modeling (e.g., Abidi and Gonzalez 1992; Clark and Yuille 1990). Consistent with this interest in the integration of multiple feature analyzers, though directed at a more abstract level, is the recent activity among statisticians who are studying methods for combining estimators for the purpose of density estimation or function approximation (e.g., Breiman 1992; Wolpert 1992).

This article reviews statistical techniques for combining multiple probability distributions. The framework with which we are concerned is that of a decision maker, henceforth referred to as the DM, who consults several experts regarding some events. The experts express their opinions to the DM in the form of probability distributions. The experts' opinions may differ due to the fact that they consider different data sets, or because they make different assumptions, or perhaps because they have different underlying theories concerning the events in question. The DM must aggregate the experts' distributions into a single distribution that can be used for decision making. This framework has been extensively studied by statisticians over the past 20 years, especially those working in the field of economic forecasting (e.g., Bates and Granger 1969; Clemen 1986; Genest and McConway 1990).

Although this framework is posed at a domain-independent level, we believe that it can be usefully applied to the study of many cognitive and perceptual processes, particularly the situation discussed above in which the outputs of multiple feature analyzers are combined during perception. Suppose, for example, that different portions of the brain each uses different monocular or binocular cues to judge the depth of a visual surface. Or suppose that different portions consider the same cues but make different assumptions regarding the visual environment and, thus, evaluate the cues differently. In either case, the brain must combine the judgments of these different portions in order to form an integrated percept of depth. This situation, as well as many other related situations, are increasingly being studied in the visual science literature. Dosher et al. (1986), for example, found that a weighted linear combination explained their data for combinations of binocular disparity and proximity-luminance covariance cues in conjunction with the kinetic depth effect. A second example is provided by Young et al. (1993) who found that the extraction of depth information based on object motion and texture gradients was consistent with a linear combination procedure.

There are many possible ways of combining probability distributions. Two classes of techniques are reviewed in this article. The first class is known as supra Bayesian methods (e.g., French 1985; Lindley 1985). The philosophy underlying this approach is that, from the viewpoint of the
Experts' Probability Assessments 869

DM, the opinions expressed by the experts are "data." Consequently, a DM using a supra Bayesian method combines the probability distributions provided by the experts with its own prior distribution via Bayes' rule. Let

- \( \theta \) denote the quantity of interest;
- \( m \) denote the number of experts;
- \( P_i = p_i(\theta \mid H_i) \) denote expert \( i \)'s probability distribution for \( \theta \) when its knowledge is \( H_i \);
- \( p(\theta \mid H) \) denote the DM's prior probability for \( \theta \) given its knowledge \( H \);
- \( p(\theta \mid H, P_1, \ldots, P_m) \) denote the DM's posterior probability for \( \theta \) given knowledge \( H \) and the experts' opinions \( P_1, \ldots, P_m \);
- \( p(P_1, \ldots, P_m \mid \theta, H) \) denote the likelihood of the experts' opinions given \( \theta \) and \( H \).

Then Bayes' rule

\[
p(\theta \mid H, P_1, \ldots, P_m) \propto p(P_1, \ldots, P_m \mid \theta, H) p(\theta \mid H)
\]

states that the DM's posterior probability is proportional to the product of its prior probability and the likelihood of the experts' opinions.

The second class of techniques reviewed in this article is referred to as linear opinion pools (e.g., Bates and Granger 1969; Genest and McConway 1990). A DM using a linear opinion pool defines its own opinion to be the linear combination of the experts' distributions with the constraint that the resulting combination is also a distribution:

\[
p(\theta) \equiv \sum_{i=1}^{m} w_i P_i
\]

where \( w_i \) are the linear coefficients. It may be the case that the DM treats itself as one of the experts, so that one of the \( P_i \) is the DM's prior distribution for \( \theta \).

The supra Bayesian and linear opinion pool approaches may be viewed as lying at opposite ends of a continuum. Supra Bayesian techniques are theoretically well-motivated and normative. That is, in many situations, supra Bayesian techniques are what one would ideally like to use. The disadvantage of these techniques is that they may be impractical on some real-world tasks. Defining an appropriate likelihood function for the experts' opinions can involve much guesswork. Moreover, evaluating this likelihood function can be computationally expensive. Because Markov chain Monte Carlo methodology can be used to evaluate complex posterior distributions, recent research using this methodology may be useful for overcoming these limitations. Gelfand et al. (1995), for example, modeled the likelihood function for the experts' opinions as a finite mixture of Beta distributions, and used Gibbs sampling to evaluate the DM's posterior distribution. It is our opinion that, due at least in part to advances...
in Bayesian statistics, supra Bayesian techniques will play an increasingly important role in future research.

Lying at the opposite end of the continuum are linear opinion pools. These pools are currently the most popular class of aggregation methods. They have the advantage that they are relatively simple, and frequently yield useful results with a moderate amount of computation. Their disadvantage is that they often lack a solid theoretical foundation. Special circumstances in which supra Bayesian techniques and linear opinion pools yield identical results are discussed below.

This review is relatively limited in its scope. It is not concerned with methods for combining the outputs of experts when these outputs are general function approximations. For example, some researchers have recently proposed using methods based on computational learning theory, particularly the PAC framework, to combine multiple function approximations (e.g., Drucker et al. 1993). These techniques will not be covered here. Our view is that there is extra structure that can be exploited if only methods for combining probability distributions are considered. Also the review covers only two out of the many possible techniques for combining distributions. More comprehensive reviews are given by Abidi and Gonzalez (1992), Chatterjee and Chatterjee (1987), French (1985), Genest and Zidek (1986) and Xu et al. (1992).

Section 2 covers supra Bayesian methods. Emphasis will be placed on making the computations tractable by adopting certain distributional assumptions. Section 3 reviews the linear opinion pool. Several methods for selecting the coefficients will be presented. It will be shown that, under some circumstances, a linear opinion pool can be justified from a Bayesian perspective. The major difficulty with combining expert opinions is that these opinions tend to be correlated or dependent. Consequently, a theme of this paper is the need for training procedures that result in experts with relatively independent opinions (cf. Meir 1994; Jordan and Jacobs 1994) or for aggregation methods that implicitly or explicitly model the dependence among the experts (cf. Hashem 1993; LeBlanc and Tibshirani 1993; Perrone 1993; Wolpert 1992). Section 4 presents analyses that show that dependent experts are worth the same as independent experts where \( k \leq m \). In some cases, an exact value for \( k \) can be given; in other cases, lower and upper bounds can be placed on \( k \).

2 Supra Bayesian Methods

Supra Bayesian methods have been developed by Agnew (1985), French (1980, 1985), Gelfand et al. (1995), Lindley (1982, 1985, 1988), Lindley et al. (1979), Morris (1974, 1977), and Winkler (1981), among others. We adopt the approach advocated by Lindley (1985) because this approach has dominated the study of supra Bayesian methods, because it makes explicit many of the underlying issues, and because it yields simple and
intuitive results. The presentation here follows closely that of Lindley (1985, 1988) and, unless otherwise noted, the results are due to Lindley.

We start by considering the case in which there is a single expert, and the quantity of interest, \( \theta \), can only assume the value \( \theta = 1 \) (henceforth denoted by the event \( A \)) or \( \theta = 0 \) (henceforth denoted by \( \tilde{A} \)). The DM assumes that the expert’s assessment of the probability of \( A \), \( P_1 = p_1(A \mid H_1) \), should be treated as data and, thus, combines this assessment with its own prior probability, \( p(A \mid H) \), via Bayes’ rule. This combination is the posterior probability for \( A \), and Bayes’ rule assumes the form:

\[
p(A \mid H, P_1) \propto p(P_1 \mid A, H) p(A \mid H)
\]

where \( p(P_1 \mid A, H) \) is the probability assigned by the DM to \( P_1 \) given event \( A \) and knowledge \( H \). An analogous equation may be written for the posterior probability of event \( \tilde{A} \). Dividing the equation for \( A \) by the equation for \( \tilde{A} \) gives the odds form:

\[
\frac{p(A \mid H, P_1)}{p(\tilde{A} \mid H, P_1)} = \frac{p(P_1 \mid A, H) p(A \mid H)}{p(P_1 \mid \tilde{A}, H) p(\tilde{A} \mid H)}
\]

and taking logarithms yields

\[
\log(A \mid H, P_1) = \log \frac{p(P_1 \mid A, H)}{p(P_1 \mid \tilde{A}, H)} + \log(A \mid H)
\]

where

\[
\log(A \mid H) = \log \frac{p(A \mid H)}{p(\tilde{A} \mid H)}
\]

denotes the log-odds for event \( A \) given knowledge \( H \). Equation 2.3 states that the posterior log-odds is equal to the sum of two terms. One term is the prior log-odds. The other term is the logarithm of the ratio of the probabilities of \( P_1 \) when \( A \) occurs to when \( \tilde{A} \) occurs. These two probabilities are typically referred to as the likelihoods for \( A \) and for \( \tilde{A} \); the logarithm of the ratio is, therefore, called a log-likelihood ratio.

This equation is a sensible way of forming posterior probabilities. Because the DM’s posterior log-odds is equal to the sum of its prior log-odds and the expert’s log-likelihood ratio, the posterior log-odds is either bigger or smaller than the prior log-odds depending on whether the log-likelihood ratio is a positive or negative number. This ratio is a measure of how much the DM values the expert’s opinion. This measure can either be positive, negative, or equal to zero. If the DM believes that the particular value that the expert has assigned to \( P_1 \) is indicative of the event \( A \), then the ratio \( p(P_1 \mid A, H) / p(P_1 \mid \tilde{A}, H) \) is greater than one, and the logarithm of this ratio is positive. The DM’s posterior log-odds for \( A \) is, therefore, greater than its prior log-odds. The likelihood ratio is less than one when the DM believes that the stated value of \( P_1 \) is indicative of \( \tilde{A} \). In this instance, the log-likelihood ratio is negative,
and the DM's posterior log-odds for $A$ is smaller than its prior log-odds. Similar logic shows that when an expert is equally likely to provide the value $P_i$ given the events $A$ or $\bar{A}$, then the log-likelihood ratio is zero, and the DM's posterior log-odds equals its prior log-odds.

Despite the intuitive and theoretical appeal of the supra Bayesian approach, it can be impractical on some real-world tasks. As mentioned above, defining an appropriate likelihood function can be problematic, and evaluating this function once it is selected can be computationally expensive. It is common, therefore, for the DM to adopt certain distributional assumptions to ameliorate these problems.

Equation 2.3 expresses the relationship between the DM's prior and posterior log-odds in terms of the expert's probabilities. It is useful if we express this relationship in terms of the expert's log-odds. Let $q_i$ denote expert $i$'s log-odds of the event $A$:

$$q_i = \log_2(p_i(A \mid H_i) = \log_2 \frac{p_i(A \mid H_i)}{p_i(\bar{A} \mid H_i)} = \log_2 \frac{P_i}{1 - P_i}$$

Then equation 2.3 may be rewritten as

$$\log(A \mid H,q_1) = \log \frac{p(q_1 \mid A,H)}{p(q_1 \mid A,\bar{H})} + \log(A \mid H)$$

Equation 2.3 expresses the relationship between the DM's prior and posterior log-odds in terms of the expert's probabilities. It is useful if we express this relationship in terms of the expert's log-odds. Let $q_i$ denote expert $i$'s log-odds of the event $A$:

$$q_i = \log_2(p_i(A \mid H_i) = \log_2 \frac{p_i(A \mid H_i)}{p_i(\bar{A} \mid H_i)} = \log_2 \frac{P_i}{1 - P_i}$$

Then equation 2.3 may be rewritten as

$$\log(A \mid H,q_1) = \log \frac{p(q_1 \mid A,H)}{p(q_1 \mid A,\bar{H})} + \log(A \mid H)$$

A common assumption is that the two distributions for the log-odds $q_1$, one given $A$ and the other given $\bar{A}$, are both normal with the same variance:

$$p(q_1 \mid A,H) \sim N(\mu_1, \sigma^2)$$

$$p(q_1 \mid \bar{A},H) \sim N(\mu_0, \sigma^2)$$

where the subscript to $\mu$ indicates whether $\theta = 1$ (denoted by the event $A$) or $\theta = 0$ (denoted by $\bar{A}$). This allows us to rewrite equation 2.6 as

$$\log(A \mid H,q_1) = \frac{\mu_1 - \mu_0}{\sigma^2} (q_1 - \frac{\mu_1 + \mu_0}{2}) + \log(A \mid H)$$

The DM's posterior log-odds is a linear combination of its prior log-odds and the expert's log-odds. First the DM adjusts the expert's log-odds, $q_1$, by a bias term $(\mu_1 + \mu_0)/2$. The bias term equals zero when $\mu_1 = -\mu_0$, a situation referred to as the symmetric case. Next the adjusted value is multiplied by $(\mu_1 - \mu_0)/\sigma^2$ and added to the DM's prior log-odds. This linear combination is not a weighted average, however, because the DM always attaches a coefficient of one to its own value.

Typically the DM has a good opinion of the expert; it will expect that the expert produces a higher probability for the event $A$ when $A$ occurs than when it does not occur. In this case, $\mu_1 > \mu_0$. If it is expected that the expert's probability for $A$ is greater than 1/2 when $A$ occurs, and less than 1/2 when $\bar{A}$ occurs, then $\mu_1 > 0 > \mu_0$. A special case is when the DM has no prior view of its own and, thus, wants to adopt
the expert’s opinion as its posterior. This occurs when the DM’s prior log-odds equals zero, meaning that it considers the events $A$ and $\bar{A}$ to be equally probable. By setting $\mu_1 = -\mu_0$ and $(\mu_1 - \mu_0)/\sigma^2 = 1$, the DM’s posterior log-odds equals the expert’s log-odds. The opposite alternative is that the DM has a poor opinion of the expert, in which case it may be that $\mu_1 < \mu_0$. Here the DM thinks that the expert usually predicts the wrong event. This does not make the expert’s opinion less valuable; it means that the DM should expect the opposite of what the expert states.

Several examples of the use of the supra Bayesian procedure are presented in Lindley (1982). In one example, the data provided by a weather forecaster are combined with the DM’s prior distribution to generate a predictor of rain whose performance is modestly better than that of either the forecaster or the DM alone. The weather forecaster supplied probabilities of rain in a subsequent 12-hr period in Chicago from July 1972 to June 1976. The DM summarized the forecaster’s stated probabilities in the event of rain, and in the event of no rain, using normal distributions. The normal distributions were fit to the forecaster’s data using maximum likelihood estimation. The forecaster’s predictions were then combined with the DM’s prior distribution using equation 2.8. The prior probability of rain in any time period was 0.25. The analyses revealed that the normal distributions provided a good fit to the forecaster’s data, though there is some evidence of mild skewness in this data. In addition, it was shown that the forecaster had a slight tendency to underestimate the probability of rain.

The supra Bayesian approach as presented so far is easily extended to the case of multiple experts. Let $q = [q_1, \ldots, q_m]^T$ denote the vector of the experts’ log-odds for the event $A$. Bayes’ rule is as above (equation 2.6) with the vector $q$ replacing the scalar $q_1$:

\[
\log(A \mid H, q) = \log \frac{p(q \mid A, H)}{p(q \mid \bar{A}, H)} + \log(A \mid H)
\]  

(2.9)

The first term on the right-hand side involves the DM’s joint distribution for the expert log-odds $q_1, \ldots, q_m$ and, thus, includes the DM’s views concerning dependencies among the experts. The normal assumption studied above extends by using the multivariate normal distribution:

\[
p(q \mid A, H) \sim N(\mu_1, \Sigma)
\]

(2.10)

\[
p(q \mid \bar{A}, H) \sim N(\mu_0, \Sigma)
\]

where the vectors $\mu_1$ and $\mu_0$ are the $m$-dimensional means for $q$ given $A$ and $\bar{A}$, respectively, and $\Sigma$ is the $m \times m$ covariance matrix. The DM’s posterior log-odds is given by

\[
\log(A \mid H, q) = [q - \frac{1}{2}(\mu_0 + \mu_1)]^T \Sigma^{-1} (\mu_1 - \mu_0) + \log(A \mid H)
\]

(2.11)

This equation includes a bias adjustment term $(\mu_0 + \mu_1)/2$ and a coefficient $\Sigma^{-1} (\mu_1 - \mu_0)$ for the experts’ log-odds $q$. 

Supra Bayesian techniques are also applicable when there are multiple
discrete events. Suppose that there are \( n \) exclusive events \( A_1, \ldots, A_n \), and
suppose that each expert provides a probability distribution for each
event. Let \( p_{ij} = p_i(A_j \mid H) \) denote the probability assigned by expert \( i \)
to event \( A_j \), and let \( q_{ij} = \log p_{ij} \) denote the logarithm of this probability.
Then the logarithm of the DM’s posterior probability for event \( A_s \) is given by

\[
\log p(A_s \mid Q, H) = c + \log p(Q \mid A_s, H) + \log p(A_s \mid H) \tag{2.12}
\]

where \( Q \) is an \( m \times n \) matrix with element \( q_{ij} \) in the \( i \)th row and \( j \)th column,
and \( c \) is a constant (\( c \) is used as a generic symbol for a constant, though
it is not the same constant each time that it appears). The DM assumes
that for each event \( A_s \), the logarithms of the expert probabilities \( q_{ij} \) have
a multivariate normal distribution with means

\[
E(q_{ij} \mid A_s, H) = \mu_{ij}s \tag{2.13}
\]

and covariances

\[
\text{cov}(q_{ij}, q_{kl} \mid A_s, H) = \sigma_{ijkl} \tag{2.14}
\]

A consequence of this assumption is that \( \log p(Q \mid A_s, H) \), the second
term on the right-hand side of equation 2.12, is a quadratic form in the
\( q_s \) plus a normalizing term that depends only on the covariances and,
thus, not on the event \( A_s \). This normalizing term can be incorporated
into the constant \( c \), so that \( \log p(Q \mid A_s, H) \) can be written as

\[
\log p(Q \mid A_s, H) = c - \frac{1}{2} \sum_{i,j,k,l} (q_{ij} - \mu_{ij}s) \sigma_{ijkl} (q_{kl} - \mu_{kl}s) \tag{2.15}
\]

where \( \sigma_{ijkl} \) are the elements of a matrix that is inverse to that with elements \( \sigma_{ijkl} \).
The equation for computing the DM’s posterior log-probability
for event \( A_s \) (equation 2.12) may now be rewritten as

\[
\log p(A_s \mid Q, H) = c + \sum_{i,j} \beta_{ij}s q_{ij} + \alpha_s + \log p(A_s \mid H) \tag{2.16}
\]

where

\[
\beta_{ij} = \sum_{k,l} \sigma_{ijkl} \mu_{kl}s \tag{2.17}
\]

and

\[
\alpha_s = \frac{1}{2} \sum_{i,j,k,l} \mu_{ij}s \sigma_{ijkl} \mu_{kl}s \tag{2.18}
\]

That is, for the DM to use the opinions provided by the \( m \) experts about
the \( n \) events, it should form its posterior log-probabilities by linearly
combining its prior log-probabilities with the logarithms of the experts’
stated probabilities. The linear coefficients depend on the DM's assessments of the experts, expressed through the means $\mu_{ij}$ and covariances $\sigma_{ijkl}$.

Lindley (1985, 1988) provides many more details regarding the supra Bayesian procedure. For example, he notes that the assumption that the experts' log-probabilities have a normal distribution is often untenable, and discusses how, in many situations, this problem can be overcome by considering contrasts instead of log-probabilities. He also shows that the computational requirements of the supra Bayesian procedure can be considerably reduced by assuming that the means and covariances of the normal distribution that characterizes the experts' log-probabilities have a restricted structure. The reader is referred to the Lindley papers for these and other matters.

In summary, we have shown how a DM can use Bayes' rule to combine its prior beliefs with experts' probability assessments. To make the resulting computations tractable, it is often necessary to assume that either the experts' log-odds or log-probabilities are normally distributed. The framework was reviewed in the cases of a single expert, multiple experts, a single discrete event, and multiple discrete events. In all instances, it was shown that with the normality assumption the DM's posterior log-odds (or log-probabilities) is the linear combination of its prior log-odds (or log-probabilities) and those of the experts.

We have already raised concerns about the supra Bayesian approach based on issues of computational expense. Some researchers have also expressed reservations based on more theoretical matters. Note, for example, that the DM assesses the joint probability of the experts' opinions given the quantity of interest $\theta$ and its knowledge $H$, whereas each expert's opinion is based on its knowledge $H_i$. Because the DM does not know each expert's knowledge, it would appear as if the DM needs to assess the probability of each expert's knowledge to evaluate the joint probability of the experts' opinions. This assessment, however, is omitted from the supra Bayesian framework.\(^1\)

Other theoretical reservations concern the order in which particular operations are performed (cf. French 1985; Genest and Zidek 1986). For example, suppose that some objective evidence becomes available such that all experts agree on the likelihood function derived from this evidence. Consider the following two procedures for determining a final distribution: (1) each expert updates its prior belief via Bayes' rule and then the experts' opinions are combined into an aggregate distribution; (2) the aggregate distribution of the experts' prior beliefs is first formed and then this distribution is updated through Bayes' rule. An aggregation method that produces the same final distribution regardless of which procedure is followed is said to possess the property of external Bayesian-ity. This property is often considered a reasonable requirement of an

---

\(^1\)We thank an anonymous reviewer for raising this issue.
aggregation technique. Note, however, that supra Bayesian methods are not externally Bayesian.

As a second example, suppose that subsets of the original events \( A_1, \ldots, A_n \) are grouped into a new set of events \( B_1, \ldots, B_r \), where \( r < n \). For instance, let the events be defined in terms of two discrete quantities, \( X \) and \( Y \), and, using new subscript notation, let \( A_{jk} \) denote the event where \( X = j \) and \( Y = k \). Let \( B_i \) denote the event that \( X = j \) (\( Y \) may take on any value). The DM is interested in the marginal probabilities of the events \( B_i \), but the experts provide probabilities for the original events \( A_{jk} \). There are at least two procedures that the DM can use to obtain the marginal probabilities: (1) compute the marginals for each expert by \( p_i(B_i) = \sum_k p_i(A_{jk}) \) and then combine the results; (2) combine the experts' probabilities about \( A_{jk} \) to obtain \( p(A_{jk} | Q, H) \) and then compute the marginal probabilities by \( p(B_i) = \sum_k p(A_{jk} | Q, H) \). If an aggregation method produces the same final marginal distribution regardless of which procedure is followed, it is said to possess the marginalization property. This property is often considered a desirable property of an aggregation technique, but it is not characteristic of supra Bayesian methods. Lindley (1985, 1988) argued that neither external Bayesianity nor the marginalization property is a reasonable requirement and, therefore, it is of no consequence that supra Bayesian techniques do not possess these features. Interestingly, McConway (1981) has shown that the only aggregation technique to possess the marginalization property is the linear opinion pool.

3 Linear Opinion Pools

This section reviews the use of linear opinion pools. When using this aggregation procedure, the DM defines its own opinion to be the linear combination of the experts' distributions with the constraint that the resulting combination is also a distribution:

\[
p(\theta) = \sum_{i=1}^{m} w_i P_i
\]

where \( w_i \) are linear coefficients or weights. A necessary condition to meet the constraint is that the weights sum to one. It is also often assumed that the weights are nonnegative (see Lawson and Hanson (1974) for the solution to least squares problems with nonnegativity constraints). The focus of this section is on methods for assigning values to the weights. Two cases are considered: weights as veridical probabilities and minimum error weights.

Weights as veridical probabilities—The DM adopts the veridical assumption when it assumes that the quantity of interest, \( \theta \), is generated by one of the experts' distributions \( P_1, \ldots, P_m \), though it is uncertain as to which one. The weight \( w_i \) is the probability that \( P_i \) is the "true" distribution and, thus, the linear opinion pool gives the marginal distribution for \( \theta \).
Experts' Probability Assessments

Statistical models known as mixture models typically adopt the veridical assumption. For example, within the artificial neural network literature, the veridical assumption is used in the mixtures-of-experts (ME) architecture proposed by Jacobs et al. (1991). It is assumed that each data item is generated as follows: given an input, one of several processes is selected from a conditional multinomial probability distribution (the distribution is conditioned on the input); the quantity of interest is sampled from the conditional distribution associated with the chosen process (again, the distribution is conditioned on the input). The ME architecture is a multinetwork architecture consisting of a "gating" network and several "expert" networks. It uses the gating network to learn the multinomial distribution, and the different expert networks to learn the distributions associated with the different processes.

The output of the architecture is the linear combination of the outputs of the expert networks. The gating network's outputs, which are non-negative and sum to one, serve as the weights. The architecture's training procedure combines aspects of associative and competitive learning. This procedure adjusts the parameters of the gating network so that, for a given input, the ith weight tends toward the probability that the distribution produced by the ith expert network is the true distribution. The training of the experts is as follows. Given an input, each expert produces an estimate of the true distribution of the quantity of interest. The expert whose estimate most closely matches the true distribution is called the winner of the competition; all other experts are called losers. It is assumed that one, and only one, expert's estimate can match the true distribution. Each expert updates its parameters in proportion to its relative performance in the competition. The overall effect is that the experts adaptively partition the data set so that each expert tends to closely approximate the true distribution for a restricted set of inputs, and different experts tend to learn the true distribution for different input sets. Because different experts receive training information on different sets of inputs, the experts' outputs are relatively independent. The veridical assumption is, therefore, often appropriate in this context.

One example of the use of the veridical assumption is provided by Nowlan (1990) who trained a mixtures-of-experts architecture on a vowel classification task. The data consisted of the first two formant values for 10 vowels spoken by 75 different speakers (Peterson and Barney 1952). In one set of simulations, 20 expert networks and a single gating network comprised the architecture; all networks received the formant values as inputs. During the course of training, different expert networks became specialized for classifying different sets of vowels. Nowlan (1990) suggested that the nature of the task decomposition discovered by the architecture is related to the positions of the vocal articulators for each of the vowel utterances; in one instance, for example, an expert network became specialized for distinguishing among the set of vowels that is spoken with the tongue toward the front of the mouth.
Despite the successes of statistical mixture models, there exist many circumstances requiring the combination of multiple experts' opinions in which it is unrealistic to assume that the experts' opinions are independent, and that the veridical assumption is valid. The experts may be nonadaptive, in which case they may come to the situation in which their opinions need to be aggregated with correlated outputs. Alternatively, the experts may be adaptive, but have opinions that are dependent because they have similar biases or receive correlated training information. For these reasons, other methods for choosing a linear opinion pool's weights have been studied.

Minimum error weights—Several researchers have proposed that the weights of the linear opinion pool be selected by performing regression on the probability of $H$ against the expert opinions $P_1, \ldots, P_m$. Two such methods, referred to as constrained and unconstrained regression, are commonly used. The linear coefficients are constrained to sum to one when using constrained regression; unconstrained regression places no constraint on the sum of the coefficients, and it employs an intercept term. Although these methods are useful for combining probability distributions, they have a broader scope of applicability in the sense that they can be used to combine any type of function approximations. Here we deviate from our practice of only considering the aggregation of probability distributions. Instead we consider the case in which experts provide point estimates of an uncertain quantity $H$ conditional on some independent variables. The DM pools these expert opinions into an aggregate point estimate. First we present the two regression methods. Next we show how each method can be justified from a Bayesian perspective in some circumstances.

Let $f_i$ denote the function that gives expert $i$'s point estimate of $H$. Suppose that the DM believes that the experts' opinions are unbiased, meaning that $E(f_i - H) = 0$. The DM may form an unbiased, minimum variance estimate of $H$, denoted $f$, by taking a weighted average of the expert opinions (Bates and Granger 1969):

$$ f = \sum_{i=1}^{m} w_if_i $$

(3.2)

The weights $w_i$ must be selected so that they minimize the variance of $f$, and so that they sum to one. A consequence of this weight selection is that the variance of the DM's estimate is guaranteed to be less than or equal to the variance of each expert's estimate. Because an unbiased estimator's variance equals its expected squared error, this means that the DM's estimate is as good as or better than any of the experts' estimates (Dickenson 1973, 1975; Perrone 1993).

Dickenson used Lagrange optimization to find the optimal weights. Suppose that the expert errors $H - f_i$ are normally distributed with zero means, and let $\sigma_y$ denote the covariance between expert $i$'s and expert $j$'s
errors. The weights are found by minimizing the objective function

\[ J = \sum_{i=1}^{m} \sum_{j=1}^{m} w_i w_j \sigma_{ij} + \lambda \left( \sum_{i=1}^{m} w_i - 1 \right) \] (3.3)

The first term on the right-hand side is the variance or expected squared error of \( f \). The second term gives the constraint that the weights sum to one. The solution to this optimization problem is

\[ w = \Sigma^{-1} I (I^T \Sigma^{-1} I)^{-1} \] (3.4)

where \( w = (w_1, \ldots, w_m) \) is the vector of weights, \( \Sigma \) is the covariance matrix for the experts' errors, and \( I \) is a vector whose elements are equal to one. If, for example, the DM believes that the experts' errors are independent, then the \( i \)th weight is proportional to the \( i \)th expert's precision \( 1/\text{var}(f_i) \).

Disadvantages of this constrained regression procedure can be illustrated by considering the case of two experts (Granger and Ramanathan 1984). The DM's estimate is

\[ f = w_1 f_1 + (1 - w_1) f_2 \] (3.5)

which can be rewritten as

\[ \theta - f_2 = w (f_1 - f_2) + \epsilon \] (3.6)

where \( \epsilon = \theta - f \) is the error. The weight \( w \) is chosen so as to minimize the expected squared error. A drawback of this procedure is that although the error \( \epsilon \) is uncorrelated with the difference \( f_1 - f_2 \), it is not necessarily uncorrelated with the individual expert estimates \( f_1 \) and \( f_2 \). It is, therefore, possible to estimate the error from the expert estimates. In this sense, the constrained regression procedure is not optimal. One possibility is to remove the constraint that the weights sum to one, but then it would no longer be the case that the DM's estimate is unbiased so long as the experts' estimates are unbiased.

An alternative is to include an additional unbiased estimate of \( \theta \), namely its unconditional mean \( E(\theta) \). In this case, the quantity \( \theta \) is given by

\[ \theta = w_1 f_1 + w_2 f_2 + w_3 E(\theta) + \epsilon \] (3.7)

where \( w_1 + w_2 + w_3 = 1 \). The weights are chosen via least-squares regression in which \( w_3 E(\theta) \) is a constant and \( w_1 \) and \( w_2 \) are unconstrained. Because \( w_3 E(\theta) \) is a constant, the DM's estimate is a linear combination of the expert estimates plus an intercept term. The error \( \epsilon \) is uncorrelated with the expert estimates.

Granger and Ramanathan (1984) advocated the use of an intercept term and the removal of the constraint that the weights sum to one. That is, the DM's estimate should be a linear combination of the expert
estimates plus an intercept term, and the weights should be selected via unconstrained least-squares regression. This method has the advantage that it yields an unbiased pooled estimate even if the expert estimates are biased.

Researchers have debated the relative merits of the constrained regression (no intercept term, weights sum to one) and unconstrained regression (intercept term, no constraint on the weights) procedures. It is clear that the unconstrained method has more "degrees-of-freedom" and, thus, will achieve a smaller sum of squared error on a set of training items (Granger and Ramanathan 1984). Nonetheless, due to possible overfitting of the training data, it is uncertain which procedure will perform better on novel data (Clemen 1986).

Meir (1994) quantified the bias and variance of linear opinion pools in the case of linear least-squares regression. Of greatest interest for our purposes is that he studied the situation where the data set is partitioned into disjoint subsets, and a different subset is used to train each expert. As compared to the case in which all experts are trained on the full data set, this training scheme can, in many situations, lead to a linear opinion pool with good performance because it results in a large decrease in the pool's variance due to the independence of the experts' opinions. This decrease tends to more than offset the concomitant increase in the pool's bias.

Bordley (1982, 1986) has shown that the constrained and unconstrained regression methods can be deduced from a Bayesian approach. In the case of constrained regression, Bordley (1982) assumed that (1) the DM's prior distribution on $\theta$ is diffuse (that is, any value of $\theta$ is equally likely); (2) the DM considers the expert errors $\theta - f_i$ to be normally distributed with mean zero and covariance matrix $\Sigma$; and (3) the expert errors are uncorrelated with the DM's prior estimates of $\theta$. Using these assumptions, $p(\theta | f_1, \ldots, f_m)$, the DM's posterior distribution on $\theta$, is a multivariate normal whose mean is given by the linear opinion pool without an intercept term and whose weights sum to one. The weights are the optimal weights selected via Lagrange optimization as described above (equation 3.4).

In the case of unconstrained regression, Bordley (1986) replaced assumptions (1) and (3) with the assumptions that the DM's prior distribution on $\theta$ is normal, and that the expert errors are correlated with the DM's prior estimates. Using these assumptions, the mean of the DM's posterior distribution is given by a linear opinion pool with an intercept term, and with no constraints on the weights. The weights are selected via unconstrained least-squares regression. When the experts' estimates are biased, the DM's estimate may be written in the form

$$f = E(\theta) + \sum_{i=1}^{m} w_i [f_i - E(f_i)]$$  (3.8)

where $E(f_i)$ is the DM's expected value for expert $i$'s estimate. That is, the DM computes its posterior expectations by adjusting its prior expecta-
Experts' Probability Assessments

...ions, either positively or negatively, in proportion to the degree to which the experts' estimates deviate from what it had expected their values to be. This is reasonable in the sense that if the experts' estimates are what the DM expects them to be, then the DM has not gained any information, and its posterior expectation equals its prior expectation.

Perrone (1993) presented several examples of the use of linear opinion pools when the experts are neural networks. One set of simulations compared different classifiers on a face recognition task. The database consisted of images of 16 human male faces. Different images of the faces were generated under various lighting conditions and with a variety of locations and orientations of the faces. During the first stage of training, 10 neural networks were individually trained to classify the faces. The networks had identical architectures, though they differed in the initial values of their weights. The outputs of the networks were combined into a linear opinion pool during a second stage of training using constrained least-squares regression. It was found that the linear opinion pool significantly outperformed all of the individual networks that comprised the pool on a novel set of images. Additional empirical results can be found in Hashem (1993) and Perrone (1993).

Recently, researchers have proposed combining linear opinion pools based on constrained or unconstrained least-squares regression with model selection techniques to achieve systems with good generalization properties (e.g., Breiman 1992; LeBlanc and Tibshirani 1993; Wolpert 1992). This combination is a special case of what Wolpert (1992) referred to as stacked generalization. To illustrate the combination, we contrast it with leave-one-out cross-validation, a common model selection procedure. Let \{x_k, \theta_k\} be a set of input-output data items, and let \(f_i^{(-k)}(x_k)\) denote the output of the ith expert in response to the input \(x_k\), when this expert has been trained using all data except data item \(k\). The prediction error for expert \(i\) is defined as

\[
PE(i) = \sum_k \left[ \theta_k - f_i^{(-k)}(x_k) \right]^2 \tag{3.9}
\]

Leave-one-out cross-validation is a "winner-take-all" model selection scheme that selects the expert with the smallest prediction error. In contrast, the combination of linear opinion pools with leave-one-out cross-validation defines the prediction error in terms of the linear aggregation of the experts' outputs:

\[
PE = \sum_k \left[ \theta_k - \sum_i w_i f_i^{(-k)}(x_k) \right]^2 \tag{3.10}
\]

The weights \(w_i\) are selected to minimize the prediction error via constrained or unconstrained regression.

Breiman (1992) compared stacked generalization with a variety of conventional statistical techniques on a wide range of linear regression tasks.
The target functions had 40 input variables and one output variable. In the first stage of the simulations, a set of experts was formed. Each expert was a linear model, and the experts differed because they each received a different subset of the input variables. The outputs of the experts were aggregated in a second stage of the simulations via a least-squares procedure that did not contain an intercept term, and that was constrained so that all the coefficients were nonnegative. The performance of this system was superior to the performances of three regression methods based upon cross-validation methodology. Additional empirical and theoretical results regarding stacked generalization can be found in Breiman (1992), LeBlanc and Tibshirani (1993), and Wolpert (1992).

In summary, we have reviewed methods that a DM can use to linearly combine experts’ probability assessments. Two cases were considered: weights as veridical probabilities and minimum error weights. In practice, linear opinion pools have proven popular because they often yield useful results with a moderate amount of computation. Objections to their use have been raised, however, based on theoretical grounds. To give just one example, it has been argued that the DM should combine the experts’ opinions in such a way as to preserve any form of expert agreement regarding the independence of the events in question (Genest and Wagner 1987). That is, it should be the case that

\[ p(A \cap B \mid P_1, \ldots, P_m) = p(A \mid P_1, \ldots, P_m) p(B \mid P_1, \ldots, P_m) \]

whenever it is each expert’s belief that \( p_i(A \cap B) = p_i(A) p_i(B) \) \( \forall i \) for events \( A \) and \( B \). This property is referred to as the independence preservation property. Note that it is not possessed by linear opinion pools except when a single expert has a weight of one and all other experts have a weight of zero, a situation referred to as a dictatorship. Genest and Wagner (1987) argued, however, that the independence preservation property is not a reasonable requirement of an aggregation procedure.

4 Value of Information from Dependent Experts

The major feature that makes the aggregation of expert opinions difficult is the high correlation or dependence that typically occurs among these opinions. This problem was alluded to in the preceding discussion; it is explicitly studied in this section where we review the results of Clemen and Winkler (1985). These authors showed that, given certain assumptions, \( m \) dependent experts are worth the same as \( k \) independent experts where \( k \leq m \). In some cases, an exact value for \( k \) can be given; in other cases, lower and upper bounds can be placed on \( k \).

Clemen and Winkler (1985) assumed that the experts provide point estimates \( f_i \) of the uncertain quantity \( \theta \), and that these estimates are unbiased, meaning that \( E(f_i - \theta) = 0 \). The vector \( \epsilon = (\epsilon_1, \ldots, \epsilon_m) \) denotes the experts’ errors where \( \epsilon_i = f_i - \theta \). It is assumed that the joint probability of the experts’ errors, \( p(\epsilon \mid \theta) \), is normally distributed with mean
Experts' Probability Assessments

zero and covariance matrix $\Sigma$. The DM's prior distribution for $\theta$, $p(\theta)$, is a normal distribution with mean $\mu_0$ and variance $\sigma_0^2$. It is assumed that the DM's prior estimation error $\mu_0 - \theta$ is uncorrelated with any of the experts' errors. Using Bayes' rule, the DM's posterior distribution is given by

$$p(\theta | f_1, \ldots, f_m) \propto p(\epsilon | \theta) p(\theta)$$

This distribution is normal with mean

$$\mu_* = (\sigma_0^{-2} \mu_0 + \mathbf{I}^T \Sigma^{-1} \mathbf{f}) \sigma_*^2$$

and variance

$$\sigma_*^2 = (\sigma_0^{-2} + \mathbf{I}^T \Sigma^{-1} \mathbf{I})^{-1}$$

where $\mathbf{I}$ is a vector whose elements are equal to one, and $\mathbf{f} = (f_1, \ldots, f_m)$ is the vector of expert opinions. Much of the analysis given below uses the fact that the posterior mean $\mu_*$ is a weighted average of the prior mean and the experts' estimates, and that the weights depend on the covariance matrix $\Sigma$. The posterior variance $\sigma_*^2$ also depends on $\Sigma$.

Suppose that the expert errors are independent, and that each expert has an error variance of $\sigma^2$. As a matter of notation, we use $m$ to denote the number of experts when these experts are dependent, and $k$ to denote the number when they are independent. The DM's prior variance can be written in the form $\sigma_0^2 = \sigma^2/k_0$. The posterior variance is then

$$\sigma_*^2 = \frac{\sigma^2}{(k_0 + k)} = (\sigma_0^{-2} + k\sigma^{-2})^{-1}$$

By comparing equations 4.3 and 4.4, we can determine the number of independent experts with error variance $\sigma^2$ that yields the same posterior variance as $m$ dependent experts with covariance matrix $\Sigma$. This number, denoted $k(\sigma^2, \Sigma)$ can be written

$$k(\sigma^2, \Sigma) = \sigma^2 \mathbf{I}^T \Sigma^{-1} \mathbf{I}$$

That is, under the given assumptions, $k(\sigma^2, \Sigma)$ independent experts are equivalent to $m$ dependent experts.

Clemen and Winkler (1985) considered three cases. The first case assumes that $\Sigma$ is an intraclass correlation matrix, meaning that all expert variances are equal and all correlations are equal. The common variance and correlation are denoted $\sigma^2$ and $\rho$, with $\rho > 0$. The inverse of the covariance matrix, $\Sigma^{-1}$, takes a relatively simple form under these conditions, and the equivalent number of independent experts is

$$k(\sigma^2, \Sigma) = m[1 + (m - 1)\rho]^{-1}$$

If $m > 1$, then $k(\sigma^2, \Sigma) < m$, meaning that positive dependence among the experts' errors reduces the information value of their opinions. The stronger the dependence, the greater is the reduction because $dk(\sigma^2, \Sigma)/d\rho$
The equivalent number of independent experts is a concave function of \( m \), whose limit may be written as

\[
\lim_{m \to \infty} k(\sigma^2, \Sigma) = \rho^{-1}
\]  

(4.7)

In other words, there is an upper limit on the number of equivalent independent experts, and on the precision of the information that can be attained by consulting dependent experts. This limit is surprisingly low. For example, if \( \rho = 0.8 \), then \( k(\sigma^2, \Sigma) = 1.25 \). After the first expert, who is worth one independent expert, all other experts combined are worth only one-fourth of an independent expert. As a second example, if \( \rho = 0.25 \), then \( k(\sigma^2, \Sigma) = 4.0 \). Consulting an infinite number of dependent experts, in this case, is no better than consulting four independent experts.

The second case considered by Clemen and Winkler (1985) assumes that the correlations among the expert errors are positive and equal, but that the error variances may differ. It is also assumed that the weights used to compute the DM's posterior mean are positive (equation 4.2). Define \( \sigma_M^2 \) and \( \sigma_m^2 \) as follows:

\[
\sigma_M^2 = \max_i \{\sigma^2_i\}; \quad \sigma_m^2 = \min_i \{\sigma^2_i\}
\]  

(4.8)

where \( \sigma^2_i \) is expert \( i \)'s variance. Then it can be shown that

\[
k(\sigma^2, \Sigma_M) < k(\sigma^2, \Sigma) < k(\sigma^2, \Sigma_m)
\]  

(4.9)

where \( \Sigma_M \) and \( \Sigma_m \) have intraclass correlation structure with correlation \( \rho \) and variances \( \sigma_M^2 \) and \( \sigma_m^2 \), respectively. In other words, an increase in the expert variances leads to a decrease in the equivalent number of independent experts.

In the general case, both the correlations among the expert errors and the expert variances may vary. As before, assume that the weights used to compute the DM's posterior mean are positive. Define

\[
\rho_R = \max_{i,j} \{\rho_{ij}\}; \quad \rho_r = \min_{i,j} \{\rho_{ij}\}; \quad \rho_0 = 0
\]  

(4.10)

where \( \rho_{ij} \) is the correlation between expert \( i \)'s and expert \( j \)'s errors. Let \( \Sigma_R \) have common correlation \( \rho_R \), \( \Sigma_r \) have common correlation \( \rho_r \), and \( \Sigma_0 \) have common correlation \( \rho_0 \), with variances equal to those in \( \Sigma \). It can be shown that

\[
k(\sigma^2, \Sigma_R) < k(\sigma^2, \Sigma) < k(\sigma^2, \Sigma_r) < k(\sigma^2, \Sigma_0)
\]  

(4.11)

An increase in the correlation among expert errors leads to a decrease in the equivalent number of independent experts.

As Clemen and Winkler (1985) pointed out, dependence among the experts can occasionally be helpful. For example, suppose that the expert errors have a common variance and correlation, but that the correlation
is negative \((-m^{-1} \leq \rho < 0\); the lower bound is necessary to make \(\Sigma\) positive definite). Then

\[ m < k(\sigma^2, \Sigma) < m^2 \]  

(4.12)

where \(m > 1\). Negative dependence can, therefore, lead to increases in the number of equivalent independent experts. As a second example, when one expert has a very large variance, it may be useful to include additional experts whose errors have high positive correlations with those of the first expert, but with smaller variances. Despite these examples, Clemen and Winkler concluded that it will generally be advantageous to include experts that are believed not to be highly correlated with each other or with the prior information, even if this means using experts with relatively high variance.

In conclusion, it appears to be the case that different aggregation procedures are appropriate for different situations. The simplest circumstance occurs when the experts' errors are uncorrelated. This may occur if it is possible to train different experts on independent data sets (Meir 1994). Alternatively, this situation may be approximated if the experts are competitive; the experts learn different mappings because they adaptively partition the data set (Jacobs et al. 1991). Aggregation is more complicated when the experts' opinions are dependent. In this case, it may be necessary for the DM to model the dependencies among the experts to achieve good performance. Two classes of aggregation procedures were reviewed in this article. With problems of easy or moderate difficulty, a “quick and dirty” procedure such as the linear opinion pool may be sufficient. A more computationally intensive technique, such as a supra Bayesian method, may be necessary for problems of greater complexity.

Recent years have seen a large increase in the number of studies on how people combine multiple sources of information, particularly in the context of visual perception. Linear opinion pools are often used to model people's perceptual cue aggregations, though some experimental results suggest that such pools are not always perfectly suited for this role. Modifications of these pools are, therefore, often explored. For example, Young et al. (1993), in their study of how people combine object motion and texture gradient visual cues to extract depth information, proposed two modifications to the basic linear opinion pool. One modification, called cue promotion, involves the use of one visual cue to provide missing information required by another cue to yield accurate perceptual judgments. The second modification is that the coefficients of the linear opinion pool may change with the visual environment, a technique referred to as dynamic reweighting. Unfortunately, there is currently no well-articulated theory to guide researchers in the selection of a model that is well-suited for the circumstance that they study. That is, it is not known which perceptual or cognitive phenomenon are best modeled using a linear opinion pool, which phenomenon should be characterized using a modified linear opinion pool, or which phenomenon
requires a more complex model such as a supra Bayesian model. This will surely be a topic of many future studies.

Acknowledgments

This work was supported in part by NIH grant MR-54770.

References


Hashem, S. 1993. *Optimal Linear Combinations of Neural Networks.* Tech. Rep. SMS 94-4, School of Industrial Engineering, Purdue University.


Received March 29, 1994; accepted March 3, 1995.
This article has been cited by:


8. Theo Gevers, Arjan Gijzenij, Joost van de Weijer, Jan-Mark Geusebroek. References 341-362. [CrossRef]


10. Leila Douha, Nabil Benoudjit, Farid Melgani. 2012. A robust regression approach for spectrophotometric signal analysis. *Journal of Chemometrics* n/a-n/a. [CrossRef]

11. MATTEO RE, GIORGIO VALENTINI. *Ensemble Methods* 20124949, . [CrossRef]


13. Haibo He. *Ensemble Learning* 108-139. [CrossRef]


23. N Benoudjitt, F Melgani, H Bouzgou. 2009. Multiple regression systems for spectrophotometric data analysis. *Chemometrics and Intelligent Laboratory Systems* 95:2, 144-149. [CrossRef]


43. Peter DayanReinforcement Learning . [CrossRef]


50. D. Tax. 2000. Combining multiple classifiers by averaging or by multiplying?. *Pattern Recognition* 33:9, 1475-1485. [CrossRef]


52. KE CHEN, HUISHENG CHI. 1999. A MODULAR NEURAL NETWORK ARCHITECTURE FOR PATTERN CLASSIFICATION BASED ON DIFFERENT FEATURE SETS. *International Journal of Neural Systems* 09:06, 563-581. [CrossRef]


