Monograph:
Marriott, S. and Harrison, R.F. (1997) The Use of Posterior Knowledge in Statistical Pattern Recognition with Particular Application to Fault Diagnosis. Research Report. ACSE Research Report 676 . Department of Automatic Control and Systems Engineering

The use of posterior knowledge in statistical pattern recognition with particular application to fault diagnosis.

S. Marriott and R. F. Harrison.

Department of Automatic Control and Systems Engineering
The University of Sheffield
Mappin Street
Sheffield, S1 3JD, U.K.

Research Report No. 676

May 1997

The use of posterior knowledge in statistical pattern recognition with particular application to fault diagnosis.

S. Marriott and R. F. Harrison.

Department of Automatic Control and Systems Engineering
The University of Sheffield
Mappin Street
Sheffield, S1 3JD, U.K.

Abstract

This paper considers the problem of posterior knowledge inclusion into fault diagnosis systems. The problem is framed in the context of set theory and elementary probability theory. A methodology of posterior probability updating is proposed for cases where fault conditions are rejected on the basis of external knowledge. Cases of exclusive, conditionally independent and dependent faults classes are considered. A possible fault hierarchy is generated following the estimation of fault class probability functions. It is shown that a simple renormalisation of existing probabilities does not apply in the dependent class case and can lead to erroneous results; the fault hierarchy may change following the exclusion of fault classes known not to have occurred. A radial basis function network with second-order regularisation is proposed as a solution to the underlying probability function estimation problem.

Key Words: Condition Monitoring, Posterior Knowledge, Fault Diagnosis, Radial Basis Function Networks.

1. The Problem

In condition monitoring applications, it is not uncommon for more than one condition type to be indicated at any one time. Hence the need for m from n classification or status indication systems where each fault condition is associated with a class, as opposed to 1 from n classification systems which are capable of indicating a single condition only. Furthermore, any predictions are tentative and may need revising when *posterior knowledge* becomes available about the true outcome, following a prediction. Posterior knowledge is knowledge about the outcome supplied by an operator, or some other source which is not available to the predictive system at the time of prediction. It is new evidence about the posterior probabilities which have been predicted for the current classification in the form of an updated output classification and differs from the new evidence about the state of the system which is typically encountered in sequential decision theory.

Any automated system which makes predictions about fault conditions will require the ability to revise probability estimates when supplied with *posterior* knowledge. How can such knowledge be incorporated so that all estimated probabilities are revised

immediately? This paper considers the problem from first principles to give an indication of possible steps towards an architecture capable of automating the process.

The condition monitoring problem involves the estimation problem in that various probability density functions are assumed to be known and must, in practice have been estimated. The inclusion of *posterior* knowledge will require these density functions to update the posterior probabilities.

Section 2 introduces the basic ideas behind the particular formulation of posterior knowledge inclusion presented in this paper and continues the discussion of relevant set theory introduced in the remainder of this section. The ideas illustrated in section 2 are expanded upon and formalised throughout the paper.

Section 3 discusses briefly general ideas concerning possible causal relationships in fault diagnosis systems. Sections 4 to 7 expand and develop earlier ideas towards a more general problem framework.

The discrete case examples of section 8 illustrate the theoretical results of previous sections. Bayes' theorem is discussed in section 9. Sections 10 to 14 consider the three cases of exclusive, independent and dependent faults respectively and derives results of relevance to the calculation of updated posterior probabilities. The hierarchy of fault possibilities is discussed in section 13.

The posterior probability function estimation problem is discussed in Sections 18 and 19 where results are derived indicating that an explosion in the number of terms (hence network outputs) is the worst-case scenario.

A result is derived in Section 20 showing that partitioning the input space into mutually exclusive events is a valid way of simplifying the estimation problem. The radial basis function network is applied to the estimation problem in Section 21.

The *input space* or *sample space* (e.g. Grimmet and Stirzaker,1992) may be divided into N, possibly overlapping, classes given by $U = C_1 \cup C_2 \ldots, \cup C_N$. This space is exhaustive and contains all possible outcomes or conditions A four class example is represented by a Venn diagram in Figure 1.

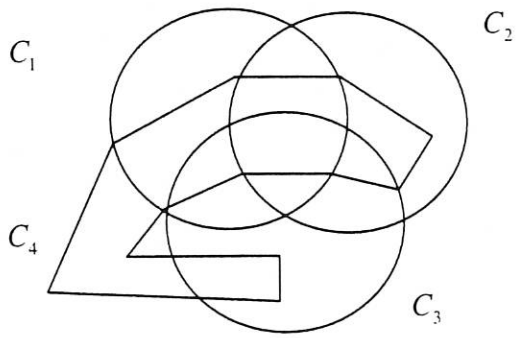

Figure 1. Abstract representation of a four class problem showing the maximum number of overlapped regions. Note that some of the possible regions of overlap may contain no members and, thus, would not exist.

Venn diagrams provide a useful way of representing the probabilities involved in updating class predictions. Figure 2 represents schematically the probability of class 1 faults occurring in a four class problem.

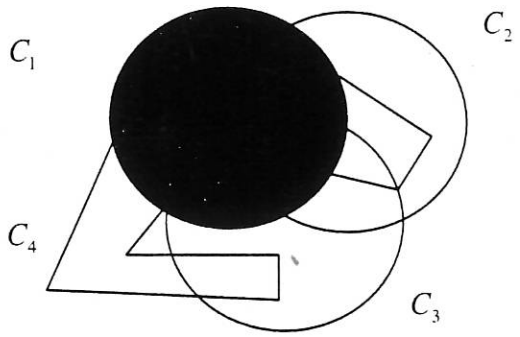

Figure 2. The shaded circle represents the total probability of class 1 occurring.

The representation of probabilities by Venn diagrams can be justified by appealing to the frequentist interpretation of probability (e.g. Kneale, 1949). Letting n_i represent the number of elements in set i (class i) and N represents the total number of elements, then $P(C_i)$ can be defined as

$$P(C_i) = \lim_{N \to \infty} \frac{n_i}{N}.$$

From this definition (e.g. Durrett, 1994) many other expressions involving the union and conjunction of sets and their respective probability definitions can be derived.

Classes may be:

i) independent
ii) dependent and exclusive
iii) dependent and non-exclusive

These three cases will be examined within this paper. Classes of type (iii) will be dealt with first being the most general—(i) and (ii) are special cases of (iii).

Independence in this context is taken to be *conditional independence* (Bernardo and Smith, 1994; Grimmet and Stirzaker, 1992). More detail is given in the relevant parts of later sections.

2. The Representation of Posterior Knowledge

There are many possible ways of representing posterior knowledge. In a probabilistic context, it is useful to represent it as a set of revised probabilities, that is, a revised probability for each class following observations of the current situation (current datum) consisting of external information. For example, a set of class posterior probabilities will be predicted for a single input datum. If it is then possible to exclude one or more classes on the basis of knowledge or reasoning not available to the

predictive system, then the list of class probabilities must be revised to give a more accurate estimation of new class posterior probabilities.

Using the definition of conditional probability for discrete events (e.g. Durrett, 1994, Grimmet and Stirzaker, 1992)

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$

where event A is conditional upon event B, the inclusion of posterior knowledge in the form of excluded classes, that is where a given system condition is known not to have occurred, can be represented by statements such as

$$P(C_1|C_2^c) = \frac{P(C_1 \cap C_2^c)}{P(C_2^c)}$$

and

$$P(C_1|C_2^c \cap C_4^c) = \frac{P(C_1 \cap C_2^c \cap C_4^c)}{P(C_2^c \cap C_4^c)}$$

the superscripted c indicates the compliment operation with respect to the *universal set*, (Grimmet and Stirzaker, 1992) thus C_2^c and C_4^c signify that classes two and four respectively have been excluded; this constitutes the new knowledge that those classes are now known not to have occurred. The revised probabilities $P(C_1|C_2^c)$ and $P(C_1|C_2^c \cap C_4^c)$ now represent the state of knowledge regarding the occurrence of class 1, after external knowledge has been incorporated, in the form of revised posterior probabilities.

Formally, the revised posterior probabilities require that the inclusion of external knowledge (evidence) be explicitly included in the notation e.g. $P(C_1|C_2^c \cap \varepsilon)$, and $P(C_1|C_2^c \cap C_4^c \cap \varepsilon)$ where the symbol 'ε' denotes the external knowledge or evidence.

Here, probabilities are required of the form $P(C_i|C_j^c \cap \varepsilon)$, $P(C_i|C_j^c \cap C_k^c \cap \varepsilon)$, and $P(C_i|C_j^c \cap C_k^c \cap C_l^c \cap \varepsilon)$ with the general form given by $P\left(C_i \middle| \bigcap_{k \in \Delta_\varepsilon} C_k^c \cap \varepsilon\right)$ where Δ_ε denotes the set of indices of the excluded classes; the exclusion being based upon external evidence.

Here, the inclusion of posterior knowledge is given in terms of classes which are known *not to have occurred* as indicated by the external knowledge. It is convenient to represent the updated posterior probabilities in terms of probabilities estimated from previous observations of system conditions, i.e. classes which have occured; these probabilities we call *positive* probabilities and they can be estimated from empirical data.

The probability of class 1 occurring given three possible classes and given the *posterior* information that class 2 has not occurred is denoted by

$$P\left(C_1 | C_2^c \cap \varepsilon\right) = \frac{P\left(C_1 \cap C_2^c\right)}{P\left(C_2^c\right)} = \frac{P\left(C_1 \cup C_2\right) - P\left(C_2\right)}{P\left(C_1 \cup C_2 \cup C_3\right) - P\left(C_2\right)}$$

using the definition of conditional probability (See Appendix A).
This situation is shown schematically in Figure 3.

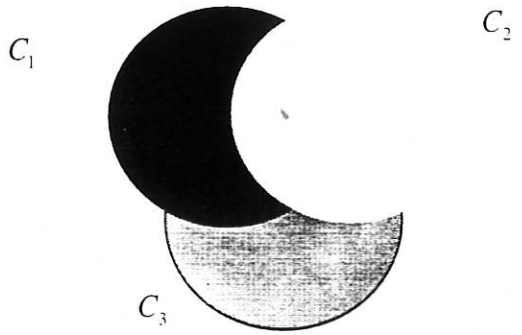

Figure 3. The diagrammatic representation of $P\left(C_1 | C_2^c \cap \varepsilon\right)$ for three dependent classes. It is the probability of the remainder of C_1 (without C_2) divided by the probability of C_1 and C_3 combined (without C_2).

The probability of class 1 occurring given a total of four possible classes and given the posterior information that class 2 has not occurred is also denoted by

$$P\left(C_1 | C_2^c \cap \varepsilon\right) = \frac{P\left(C_1 \cap C_2^c\right)}{P\left(C_2^c\right)}.$$

This is illustrated in Figure 4.

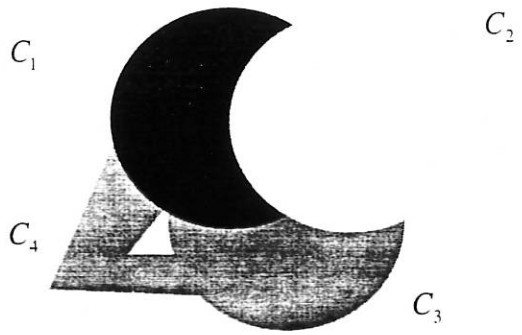

Figure 4. The diagrammatic representation of $P\left(C_1 | C_2^c \cap \varepsilon\right)$ for four dependent classes. The probability of the remainder of class 1 occurring is divided by the total remaining probability excluding class 2 to give the remaining probability of class 1 occurring.

For dependent classes, the underlying general pattern appears to consist of finding the difference of unions of those sets involved in the numerator and denominator of the conditional probability expression and finding the ratio of the respective probabilities. For example, for the four class problem, $P\left(C_1 | C_2^c \cap \varepsilon\right)$ can be written as

$$P(C_1|C_2^c \cap \varepsilon) = \frac{P(C_1 \cap C_2^c)}{P(C_2^c)} = \frac{P(C_1 \cup C_2) - P(C_2)}{P\left(\bigcup_{l=1}^{4} C_l\right) - P(C_2)}$$

where the posterior information has been included in terms of *positive* probabilities, that is, those which can be estimated directly, or computed from estimated probabilities. Similarly, the inclusion of further posterior information that class C_4 has also been excluded can be written as

$$P(C_1|C_2^c \cap C_4^c \cap \varepsilon) = \frac{P(C_1 \cap C_2^c \cap C_4^c)}{P(C_2^c \cap C_4^c)} = \frac{P(C_1 \cup C_2 \cup C_4) - P(C_2 \cup C_4)}{P\left(\bigcup_{l=1}^{4} C_l\right) - P(C_2 \cup C_4)}$$

Here, the class unions have increased by a single member (class 4) which is to be excluded to give the revised probabilities. This is shown in Figure 5.

Figure 5. A representation of $P(C_1|C_2^c \cap C_4^c \cap \varepsilon)$ where class 4 has also been excluded from Figure 4.

The set of fault classes is taken to be exhaustive with the universal set given by the union of all classes including a "no-fault" class.

3. Cause and Effect

Figure 6 shows a possible sub-system of a larger system to illustrate the ideas of cause and effect. Conditions in components A and C occur independently of those in component B and vice-versa. However, the occurrence of a condition in component A entails a condition in component C. A condition in C can also occur independently of conditions in A. Thus, there is a causal link between A and C.

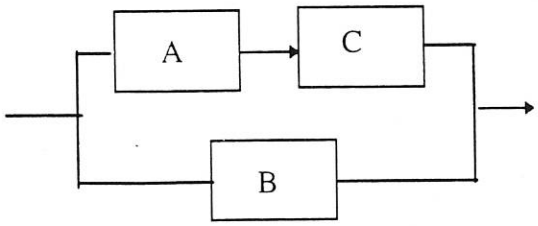

Figure 6. A schematic representation of a group of interconnected sub-systems which form part of a system model. Here, sub-systems A and C are causally connected.

Figure 7 illustrates this situation in terms of sets. For $A \subset C$, $(A \neq C)$ the condition denoted by class C will always occur when the condition denoted by class A occurs. Thus,

$$P(C|A) = \frac{P(A \cap C)}{P(A)} = \frac{P(A)}{P(A)} = 1$$

This is not necessarily the case the other way round where

$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{P(A)}{P(C)} < 1$ if class A and class C are distinct (not equal). The case that $A = C$ is excluded because two such classes would be indistinguishable.

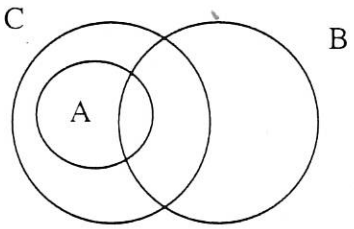

Figure 7. A Venn diagram showing the relationships between the event sets of figure 6. When class A occurs it causes class C to occur. Note that class C events do not necessarily generate a class A event.

4. The General Form of Posterior Knowledge Inclusion Where One Or More Dependent Classes Are Excluded.

For the general case, where a set of dependent classes is excluded, the following notation is introduced:, Δ_r and Δ_ε denote the index sets of included and excluded sets respectively where $\Delta_r = \{\delta_1, \delta_2, \ldots, \delta_{Nr}\}$ and $\Delta_\varepsilon = \{\delta_{Nr+1}, \delta_{Nr+2}, \ldots, \delta_N\}$ N_r is the number of included classes, N is the total number of possible classes. The delta notation is used to denote that the class indices are not necessarily selected on the basis of ordering e.g. it could be that for a five class problem, $\Delta_r = \{1,3,5\}$ and $\Delta_\varepsilon = \{2,4\}$ in which case $\delta_1 = 1$, $\delta_2 = 3$, $\delta_3 = 5$, $\delta_4 = 2$ and $\delta_5 = 4$ where classes two and four have been excluded.

In general, to calculate the updated probabilities, given *posterior* knowledge,

$$P\left(C_{\delta_i} | C^c_{\delta_{Nr+1}} \cap C^c_{\delta_{Nr+2}} \cap \ldots \cap C^c_{\delta_N} \cap \varepsilon\right) = \frac{P\left(C_{\delta_i} \cap C^c_{\delta_{Nr+1}} \cap C^c_{\delta_{Nr+2}} \cap \ldots \cap C^c_{\delta_N}\right)}{P\left(C^c_{\delta_{Nr+1}} \cap C^c_{\delta_{Nr+2}} \cap \ldots \cap C^c_{\delta_N}\right)}$$

$$= \frac{P\left(\bigcup_j C_{\delta_j}\right) - P\left(\bigcup_k C_{\delta_k}\right)}{P\left(\bigcup_l C_{\delta_l}\right) - P\left(\bigcup_k C_{\delta_k}\right)} \qquad (1)$$

where δ_i is the i th index, $\delta_i \in \{1,2,\ldots,N\}$, $j \in \{\delta_i\} \cup \Delta_\varepsilon$, $k \in \Delta_\varepsilon$, and $l \in \Delta_r \cup \Delta_\varepsilon$.

For the numerator, the probability of the union of all classes omitted is subtracted from the probability of this union augmented by the class of interest i.e. $P\left(C_{\delta_i} \cup \left(\bigcup_k C_{\delta_k}\right)\right) - P\left(\bigcup_k C_{\delta_k}\right)$. For the denominator, the probability of the union of all excluded classes is subtracted from the probability of the total number of classes i.e. the certain event. This calculation suggests that an incremental procedure is possible (see Section 8)

The proof of equation (1) is reserved until Section 7 when the dependence of the class on the data is included.

5. Union of Overlapping Sets.

Now that the general form of the updated posterior probabilities $P\left(C_{\delta_i} | C^c_{\delta_{Nr-1}} \cap C^c_{\delta_{Nr-2}} \cap \ldots \cap C^c_{\delta_N} \cap \varepsilon\right)$ is given in terms of unions of sets, the general form of $P\left(\bigcup_{s=1}^K C_{\delta_s}\right)$ is required (e.g Durrett, 1994,Grimmet and Stirzaker, 1992) where K is the number of sets involved in the union. This is expanded to give

$$\begin{aligned} P\left(\bigcup_{s=1}^K C_{\delta_s}\right) = & \sum_{i=1}^K P\left(C_{\delta_i}\right) \\ & - \sum_{i<j}^K P\left(C_{\delta_i} \cap C_{\delta_j}\right) \\ & + \sum_{i<j<k}^K P\left(C_{\delta_i} \cap C_{\delta_j} \cap C_{\delta_k}\right) \qquad (2) \\ & \vdots \\ & + (-1)^{K+1} P\left(C_{\delta_1} \cap C_{\delta_2} \cap \ldots \cap C_{\delta_K}\right) \end{aligned}$$

in terms of positive probabilities.

Recall that the δ notation is used to reflect the fact that the indices *i,j,k...* are not necessarily consecutive or ordered sequentially.

5.1 Examples:

5.1.1 Example 1

For three sets, the union is given by

$$\begin{aligned} P(C_1 \cup C_2 \cup C_3) = & P(C_1) + P(C_2) + P(C_3) \\ & - P(C_1 \cap C_2) - P(C_1 \cap C_3) - P(C_2 \cap C_3) \\ & + P(C_1 \cap C_2 \cap C_3) \end{aligned}$$

as shown in Figure 8

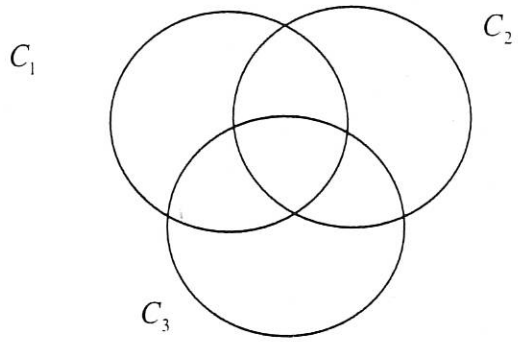

Figure 8. The union of three sets. It is calculated by subtracting all overlaps and replacing the missing "centre"

5.1.2 Example 2

For four sets we have

$$\begin{aligned}P(C_1 \cup C_2 \cup C_3 \cup C_4) = {} & P(C_1)+P(C_2)+P(C_3)+P(C_4) \\ & -P(C_1 \cap C_2)-P(C_1 \cap C_3)-P(C_1 \cap C_4) \\ & -P(C_2 \cap C_3)-P(C_2 \cap C_4)-P(C_3 \cap C_4) \\ & +P(C_1 \cap C_2 \cap C_3)+P(C_1 \cap C_2 \cap C_4) \\ & +P(C_1 \cap C_3 \cap C_4)+P(C_2 \cap C_3 \cap C_4) \\ & -P(C_1 \cap C_2 \cap C_3 \cap C_4)\end{aligned}$$

6. Posterior Probabilities In Terms of Overlapping Sets: Examples

Applying the formula for unions of overlapping sets (equation (2)) to the general formula for the inclusion of posterior knowledge (equation (1)) allows the posterior probabilities to be computed.

6.1 Examples of Posterior Knowledge Inclusion in Terms of Positive Probabilities.

6.1.1 For Three Classes where a Single Class has Been Excluded:

$$\begin{aligned}P(C_1 | C_2^c \cap \varepsilon) &= \frac{P(C_1 \cap C_2^c)}{P(C_2^c)} \\ &= \frac{P(C_1 \cup C_2)-P(C_2)}{P(C_1 \cup C_2 \cup C_3)-P(C_2)} \\ &= \frac{P(C_1)-P(C_1 \cap C_2)}{P(C_1)+P(C_3)-P(C_1 \cap C_2)-P(C_1 \cap C_3)-P(C_2 \cap C_3)+P(C_1 \cap C_2 \cap C_3)}\end{aligned}$$

6.1.2 For four classes:

$$P\left(C_1 \mid C_2^c \cap \varepsilon\right) = \frac{P\left(C_1 \cap C_2^c\right)}{P\left(C_2^c\right)}$$

$$= \frac{P\left(C_1 \cup C_2\right) - P\left(C_2\right)}{P\left(C_1 \cup C_2 \cup C_3 \cup C_4\right) - P\left(C_2\right)}$$

$$= \frac{P\left(C_1\right) - P\left(C_1 \cap C_2\right)}{\begin{aligned} & P\left(C_1\right) + P\left(C_3\right) + P\left(C_4\right) \\ & - P\left(C_1 \cap C_2\right) - P\left(C_1 \cap C_3\right) - P\left(C_1 \cap C_4\right) \\ & - P\left(C_2 \cap C_3\right) - P\left(C_2 \cap C_4\right) - P\left(C_2 \cap C_4\right) \\ & + P\left(C_1 \cap C_2 \cap C_3\right) + P\left(C_1 \cap C_2 \cap C_4\right) \\ & + P\left(C_1 \cap C_3 \cap C_4\right) + P\left(C_2 \cap C_3 \cap C_4\right) \\ & - P\left(C_1 \cap C_2 \cap C_3 \cap C_4\right) \end{aligned}}$$

6.1.3 The Exclusion of *two* Classes in the Four class Example:

$$P\left(C_1 \mid C_2^c \cap C_4^c \cap \varepsilon\right) = \frac{P\left(C_1 \cap C_2^c \cap C_4^c\right)}{P\left(C_2^c \cap C_4^c\right)}$$

$$= \frac{P\left(C_1 \cup C_2 \cup C_4\right) - P\left(C_2 \cup C_4\right)}{P\left(C_1 \cup C_2 \cup C_3 \cup C_4\right) - P\left(C_2 \cup C_4\right)}$$

$$= \frac{\begin{aligned} & P\left(C_1\right) - P\left(C_1 \cap C_2\right) \\ & - P\left(C_1 \cap C_4\right) + P\left(C_1 \cap C_2 \cap C_4\right) \end{aligned}}{\begin{aligned} & P\left(C_1\right) + P\left(C_3\right) \\ & - P\left(C_1 \cap C_2\right) - P\left(C_1 \cap C_3\right) - P\left(C_1 \cap C_4\right) \\ & - P\left(C_2 \cap C_3\right) - P\left(C_3 \cap C_4\right) \\ & + P\left(C_1 \cap C_2 \cap C_3\right) + P\left(C_1 \cap C_2 \cap C_4\right) \\ & + P\left(C_1 \cap C_3 \cap C_4\right) + P\left(C_2 \cap C_3 \cap C_4\right) \\ & - P\left(C_1 \cap C_2 \cap C_3 \cap C_4\right) \end{aligned}}$$

7. All Probabilities are Conditional Upon the Input Space

So far, the dependence of the classes on the underlying data space has been omitted. In actuality, these classes are labels attached to regions of the space. Denoting the underlying data variable by the n-dimensional vector $\mathbf{x}$, and assuming for now that the elements of $\mathbf{x}$ are discrete random variables, the expression that the updated class probabilities including posterior information can be stated as

$$P\left(C_{\delta_i} | C^c_{\delta_{Nr-1}} \cap C^c_{\delta_{Nr-2}} \cap \ldots \cap C^c_{\delta_N} \cap \mathbf{x} \cap \varepsilon\right) = \frac{P\left(C_{\delta_i} \cap C^c_{\delta_{Nr-1}} \cap C^c_{\delta_{Nr-2}} \cap \ldots \cap C^c_{\delta_N} \cap \mathbf{x}\right)}{P\left(C^c_{\delta_{Nr-1}} \cap C^c_{\delta_{Nr-2}} \cap \ldots \cap C^c_{\delta_N} \cap \mathbf{x}\right)}$$

$$= \frac{P\left(\left(\bigcup_j C_{\delta_j}\right) \cap \mathbf{x}\right) - P\left(\left(\bigcup_k C_{\delta_k}\right) \cap \mathbf{x}\right)}{P\left(\left(\bigcup_l C_{\delta_l}\right) \cap \mathbf{x}\right) - P\left(\left(\bigcup_k C_{\delta_k}\right) \cap \mathbf{x}\right)} \qquad (3)$$

where $j \in \{\delta_i\} \cup \Delta_\varepsilon$, $k \in \Delta_\varepsilon$, and $l \in \Delta_r \cup \Delta_\varepsilon$. From equation (3) we get

$$P\left(C_{\delta_i} | C^c_{\delta_{Nr-1}} \cap C^c_{\delta_{Nr-2}} \cap \ldots \cap C^c_{\delta_N} \cap \mathbf{x} \cap \varepsilon\right) = \frac{P\left(\left(\bigcup_j C_{\delta_j}\right) \cap \mathbf{x}\right) - P\left(\left(\bigcup_k C_{\delta_k}\right) \cap \mathbf{x}\right)}{P\left(\left(\bigcup_l C_{\delta_l}\right) \cap \mathbf{x}\right) - P\left(\left(\bigcup_k C_{\delta_k}\right) \cap \mathbf{x}\right)}$$

$$= \frac{P\left(\bigcup_j C_{\delta_j} \middle| \mathbf{x}\right) P(\mathbf{x}) - P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right) P(\mathbf{x})}{P\left(\bigcup_l C_{\delta_l} \middle| \mathbf{x}\right) P(\mathbf{x}) - P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right) P(\mathbf{x})}$$

$$= \frac{P\left(\bigcup_j C_{\delta_j} \middle| \mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right)}{P\left(\bigcup_l C_{\delta_l} \middle| \mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right)}$$

where the rule $P(A \cap B) = P(A|B)P(B)$ has been applied.
The above expression reduces to

$$\frac{P\left(\mathbf{x} \cap \bigcup_j C_{\delta_j}\right) - P\left(\mathbf{x} \cap \bigcup_k C_{\delta_k}\right)}{P(\mathbf{x}) - P\left(\mathbf{x} \cap \bigcup_k C_{\delta_k}\right)}$$

if *all* input vectors have class labels associated with them i.e. the set of classes is exhaustive.

The final expression for the revised probabilities

$$P\left(C_{\delta_i}\left|\bigcap_k C_{\delta_k}^c \cap \mathbf{x} \cap \varepsilon\right.\right) = \frac{P\left(\left(\bigcup_j C_{\delta_j}\right)\middle|\mathbf{x}\right) - P\left(\left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right)}{P\left(\left(\bigcup_l C_{\delta_l}\right)\middle|\mathbf{x}\right) - P\left(\left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right)} \tag{4}$$

Equation (4) can be proved formally as follows:

$$P\left(C_{\delta_i}\left|\left(\bigcap_k C_{\delta_k}^c\right) \cap \mathbf{x} \cap \varepsilon\right.\right) = \frac{P\left(C_{\delta_i} \cap \left(\bigcap_k C_{\delta_k}^c\right) \cap \mathbf{x}\right)}{P\left(\left(\bigcap_k C_{\delta_k}^c\right) \cap \mathbf{x}\right)} \text{ by definition of conditional}$$

probability

$$= \frac{P\left(C_{\delta_i} \cap \left(\bigcap_k C_{\delta_k}^c\right)\middle|\mathbf{x}\right)P(\mathbf{x})}{P\left(\left(\bigcap_k C_{\delta_k}^c\right)\middle|\mathbf{x}\right)P(\mathbf{x})} \text{ by } P(A \cap B) = P(A|B)P(B)$$

$$= \frac{P\left(C_{\delta_i} \cap \left(\bigcap_k C_{\delta_k}^c\right)\middle|\mathbf{x}\right)}{P\left(\left(\bigcap_k C_{\delta_k}^c\right)\middle|\mathbf{x}\right)} \text{ by cancellation of } P(\mathbf{x})$$

$$= \frac{P\left({}^cC_{\delta_i} \cap \left[\left(\bigcap_k C_{\delta_k}^c\right)^c\right]\middle|\mathbf{x}\right)}{P\left(U \cap \left[\left(\bigcap_k C_{\delta_k}^c\right)^c\right]^c\middle|\mathbf{x}\right)} \text{ by } \left(A^c\right)^c = A \text{ and } U \cap A = A$$

$$= \frac{P\left(C_{\delta_i} \cup \left(\bigcap_k C_{\delta_k}^c\right)^c\middle|\mathbf{x}\right) - P\left(\left(\bigcap_k C_{\delta_k}^c\right)^c\middle|\mathbf{x}\right)}{P\left(U \cup \left(\bigcap_k C_{\delta_k}^c\right)^c\middle|\mathbf{x}\right) - P\left({}^c\left(\bigcap_k C_{\delta_k}^c\right)\middle|\mathbf{x}\right)} \text{ by } P\left(A \cap B^c\right) = P(A \cup B) - P(B)$$

$$= \frac{P\left(C_{\delta_i} \cup \left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right) - P\left(\left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right)}{P\left(U \cup \left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right) - P\left(\left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right)} \text{ by de Morgan's law(e.g. Applebaum, 1996)}$$

$$= \frac{P\left(C_{\delta_i} \cup \left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right) - P\left(\left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right)}{P(U|\mathbf{x}) - P\left(\left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right)}$$

$$= \frac{P\left(\left(\bigcup_j C_{\delta_j}\right)\middle|\mathbf{x}\right) - P\left(\left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right)}{P\left(\left(\bigcup_l C_{\delta_l}\right)\middle|\mathbf{x}\right) - P\left(\left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right)}$$

where the fact that the union of the classes is exhaustive has been used.

If $\boldsymbol{x}$ is real-valued (continuous) then appropriate probability density functions of the form

$$p\left(\mathbf{x} \cap \left(\bigcup_j C_{\delta_j}\right)\right) = P\left(\left(\bigcup_j C_{\delta_j}\right)\middle|\mathbf{x}\right) p(\mathbf{x})$$

are substituted into the previous argument where $p(.)$ denotes a probability *density* function.

After including the dependency upon $\boldsymbol{x}$, equation (2) is now written in terms of probabilities conditional upon the input

$$\begin{aligned} P\left(\bigcup_{s=1}^{K} C_{\delta_s}\middle|\mathbf{x}\right) = & \sum_{i=1}^{K} P\left(C_{\delta_i}\middle|\mathbf{x}\right) \\ & - \sum_{i<j}^{K} P\left(C_{\delta_i} \cap C_{\delta_j}\middle|\mathbf{x}\right) \\ & + \sum_{i<j<k}^{K} P\left(C_{\delta_i} \cap C_{\delta_j} \cap C_{\delta_k}\middle|\mathbf{x}\right) \qquad (5) \\ & \vdots \\ & + (-1)^{K+1} P\left(C_{\delta_1} \cap C_{\delta_2} \cap \ldots \cap C_{\delta_K}\middle|\mathbf{x}\right) \end{aligned}$$

to include the conditional probabilities of equation (4). Equation (5) can be proved easily by using the distributivity of set relations and substituting $C_{\delta_i} \cap \mathbf{x}$ for C_{δ_i} in equation (2).

8. A Simple Example in Terms of Relative Frequencies.

In principle, the probability estimation problem can be solved by counting the number of occurrences of classes on a case-by-case basis. Here, the probabilities are represented by relative frequency observations. In practice, this method requires a large amount of data (Bishop, 1995) and may not be accurate thus necessitating the use of a continuous (i.e. not discrete-valued) estimator. The example presented here is intended to illustrate the principles discussed in the previous section in a simple context.

The probabilities of single classes are approximated by

$$P(C_i|\mathbf{x}) = \frac{P(C_i \cap \mathbf{x})}{P(\mathbf{x})}$$
$$= \frac{n(C_i \cap \mathbf{x}) / N(\mathbf{x})}{n(\mathbf{x}) / N(\mathbf{x})}$$
$$= \frac{n(C_i \cap \mathbf{x})}{n(\mathbf{x})}$$

where $N(\mathbf{x})$ is the total number of condition occurrences within a given region of the data space $\boldsymbol{x}$. Similarly for two classes occurring simultaneously

$$P(C_i \cap C_j|\mathbf{x}) = \frac{P(C_i \cap C_j \cap \mathbf{x})}{P(\mathbf{x})}$$
$$= \frac{n(C_i \cap C_j \cap \mathbf{x})}{n(\mathbf{x})}$$

or for all classes:

$$P(C_1 \cap C_2 \cap \ldots \cap C_N|\mathbf{x}) = \frac{P(C_1 \cap C_2 \cap \ldots \cap C_N \cap C_x)}{P(\mathbf{x})}$$
$$= \frac{n(C_1 \cap C_2 \cap \ldots \cap C_N \cap \mathbf{x})}{n(\mathbf{x})}$$

A numerical example of overlapping classes with all elements associated with $\boldsymbol{x}$ is shown in Figure 9.

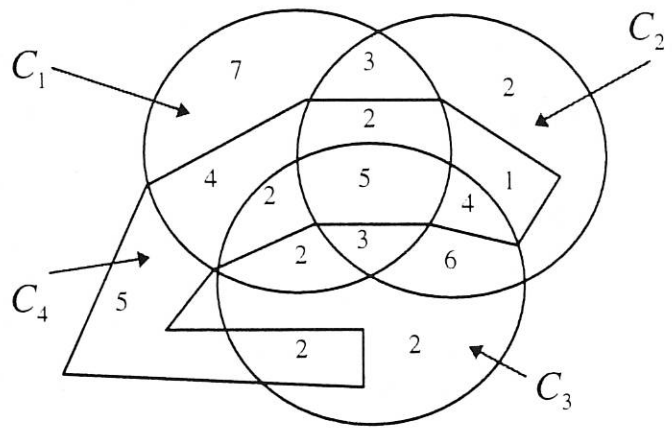

Figure 9. A diagrammatic representation of the numerical example.

The class frequency counts are given by

$$n(C_1 \cap \mathbf{x}) = 28,\ n(C_2 \cap \mathbf{x}) = 26,\ n(C_3 \cap \mathbf{x}) = 26,\ n(C_4 \cap \mathbf{x}) = 25,$$

$$n(C_1 \cap C_2 \cap \mathbf{x}) = 13,\ n(C_1 \cap C_3 \cap \mathbf{x}) = 12,\ n(C_1 \cap C_4 \cap \mathbf{x}) = 13,$$
$$n(C_2 \cap C_3 \cap \mathbf{x}) = 18,\ n(C_2 \cap C_4 \cap \mathbf{x}) = 12,\ n(C_3 \cap C_4 \cap \mathbf{x}) = 13,$$

$n(C_1 \cap C_2 \cap C_3 \cap \mathbf{x}) = 8$, $n(C_1 \cap C_2 \cap C_4 \cap \mathbf{x}) = 7$, $n(C_1 \cap C_3 \cap C_4 \cap \mathbf{x}) = 7$,
$n(C_2 \cap C_3 \cap C_4 \cap \mathbf{x}) = 9$,

$n(C_1 \cap C_2 \cap C_3 \cap C_4 \cap \mathbf{x}) = 5$

The total number of condition occurrences across the four classes is given by

$$N(\mathbf{x}) = 28+26+26+25-13-12-13-18-12-13+8+7+7+9-5$$
$$= 105-81+31-5$$
$$= 50$$

Now the relative frequencies (probabilities) of the *singleton* classes (the total probability of each class as a whole) can be calculated by

$$P(C_1|\mathbf{x}) = \frac{n(C_1 \cap \mathbf{x})}{N(\mathbf{x})} = \frac{28}{50} = 0.56$$

Similarly,

$$P(C_2|\mathbf{x}) = \frac{26}{50} = 0.52,\ P(C_3|\mathbf{x}) = \frac{26}{50} = 0.52, \text{ and } P(C_4|\mathbf{x}) = \frac{25}{50} = 0.5$$

Note that $\sum_{i=1}^{4} P(C_i) \neq 1$ because the classes are not exclusive, instead, the union of the four classes $P\left(\bigcup_{i=1}^{4} C_i\right) = 1$.

The class pairs are given by

$$P(C_1 \cap C_2|\mathbf{x}) = \frac{n(C_1 \cap C_2 \cap \mathbf{x})}{N(\mathbf{x})} = \frac{13}{50} = 0.26$$

$$P(C_1 \cap C_3|\mathbf{x}) = \frac{12}{50} = 0.24,\ P(C_1 \cap C_4|\mathbf{x}) = \frac{13}{50} = 0.26,\ P(C_2 \cap C_3) = \frac{18}{50} = 0.36$$

$$P(C_2 \cap C_4|\mathbf{x}) = \frac{12}{50} = 0.24,\ P(C_3 \cap C_4|\mathbf{x}) = \frac{13}{50} = 0.26$$

the class triples by

$$P(C_1 \cap C_2 \cap C_3|\mathbf{x}) = \frac{n(C_1 \cap C_2 \cap C_3 \cap \mathbf{x})}{N(\mathbf{x})} = \frac{8}{50} = 0.16$$

$$P(C_1 \cap C_2 \cap C_4|\mathbf{x}) = \frac{7}{50} = 0.14,\ P(C_1 \cap C_3 \cap C_4) = \frac{7}{50} = 0.14$$

$$P(C_2 \cap C_3 \cap C_4|\mathbf{x}) = \frac{9}{50} = 0.18$$

and finally, $P(C_1 \cap C_2 \cap C_3 \cap C_4|\mathbf{x}) = \frac{n(C_1 \cap C_2 \cap C_3 \cap C_4 \cap \mathbf{x})}{N(\mathbf{x})} = \frac{5}{50} = 0.1$.

The revised probabilities given the *posterior* knowledge, C_2^c, (i.e. not class 2 in this case) are given by

$$P(C_1|C_2^c \cap \mathbf{x} \cap \varepsilon) = \frac{P(C_1 \cup C_2|\mathbf{x}) - P(C_2|\mathbf{x})}{1 - P(C_2|\mathbf{x})} = \frac{P(C_1|\mathbf{x}) - P(C_1 \cap C_2|\mathbf{x})}{1 - P(C_2|\mathbf{x})}$$

$$= \frac{\frac{28}{50} - \frac{13}{50}}{\frac{50}{50} - \frac{26}{50}} = \frac{15}{24} = 0.625$$

Similarly,

$$P(C_3|C_2^c \cap \mathbf{x}\varepsilon) = \frac{\frac{26}{50} - \frac{18}{50}}{\frac{50}{50} - \frac{26}{50}} = \frac{8}{24} = 0.333$$

$$P(C_4|C_2^c \cap \mathbf{x} \cap \varepsilon) = \frac{\frac{26}{50} - \frac{13}{50}}{\frac{50}{50} - \frac{26}{50}} = \frac{13}{24} = 0.5417$$

If class 4 is also excluded, the revised probabilities become

$$P(C_1|C_2^c \cap C_4^c \cap \mathbf{x} \cap \varepsilon) = \frac{P(C_1 \cap C_2^c \cap C_4^c \cap \mathbf{x})}{P(C_2^c \cap C_4^c)}$$

$$= \frac{P(C_1 \cup C_2 \cup C_4|\mathbf{x}) - P(C_2 \cup C_4|\mathbf{x})}{P(C_1 \cup C_2 \cup C_3 \cup C_4|\mathbf{x}) - P(C_2 \cup C_4|\mathbf{x})}$$

$$= \frac{\begin{array}{l} P(C_1|\mathbf{x}) - P(C_1 \cap C_2|\mathbf{x}) \\ - P(C_1 \cap C_4|\mathbf{x}) + P(C_1 \cap C_2 \cap C_4|\mathbf{x}) \end{array}}{\begin{array}{l} P(C_1|\mathbf{x}) + P(C_3|\mathbf{x}) \\ - P(C_1 \cap C_2|\mathbf{x}) - P(C_1 \cap C_3|\mathbf{x}) - P(C_1 \cap C_4|\mathbf{x}) \\ - P(C_2 \cap C_3|\mathbf{x}) - P(C_3 \cap C_4|\mathbf{x}) \\ + P(C_1 \cap C_2 \cap C_3|\mathbf{x}) + P(C_1 \cap C_2 \cap C_4|\mathbf{x}) \\ + P(C_1 \cap C_3 \cap C_4|\mathbf{x}) + P(C_2 \cap C_3 \cap C_4|\mathbf{x}) \\ - P(C_1 \cap C_2 \cap C_3 \cap C_4|\mathbf{x}) \end{array}}$$

$$= \frac{\frac{28}{50} - \frac{13}{50} - \frac{13}{50} + \frac{7}{50}}{\frac{28}{50} + \frac{26}{50} - \frac{13}{50} - \frac{12}{50} - \frac{13}{50} - \frac{18}{50} - \frac{13}{50} + \frac{8}{50} + \frac{7}{50} + \frac{7}{50} + \frac{9}{50} - \frac{5}{50}}$$

$$= \frac{9}{11} = 0.8182$$

and

$$P\left(C_3|C_2^c \cap C_4^c \cap \mathbf{x} \cap \varepsilon\right) = \frac{\frac{26}{50} - \frac{18}{50} - \frac{13}{50} + \frac{9}{50}}{\frac{11}{50}} = \frac{4}{11} = 0.3636$$

Note that had a simple renormalisation been used following the exclusion of class 2, the results would have been given by

$$P\left(C_1|\mathbf{x}\right) = \frac{P\left(C_1|\mathbf{x}\right)}{P\left(C_1 \cup C_3|\mathbf{x}\right)} = \frac{P\left(C_1|\mathbf{x}\right)}{P\left(C_1|\mathbf{x}\right) + P\left(C_3|\mathbf{x}\right) - P\left(C_1 \cap C_3|\mathbf{x}\right)}$$

$$= \frac{\frac{28}{50}}{\frac{28}{50} + \frac{26}{50} - \frac{12}{50}} = \frac{28}{42} = 0.6666$$

Similarly,

$$P\left(C_3|\mathbf{x}\right) = \frac{26}{42} = 0.6190$$

A simple renormalisation is not sufficient because the classes are not exclusive or independent and, consequently, the exclusion of classes 2 and 4 affects the probabilities of occurrence of classes 1 and 3 depending upon the extent of 'coupling' between the respective classes (see Table 1).

Class	Adjusted for overlap	Simple Renormalisation
1	0.56	0.6666
3	0.52	0.6190

Table 1 The effects of adjusting the new posterior probabilities by taking into account the overlaps with the excluded classes. Note that the probabilities obtained using a simple renormalisation are only valid for exclusive or independent classes.

9. The Use of Bayes Theorem.

As will be discussed in sections 18 and 19, posterior probabilities can be estimated directly if certain techniques are used. In some cases, however, it may be more appropriate to use Bayesian decision theory, and compute the posterior probabilities indirectly rather than directly estimating them. Bayesian decision theory is a framework for calculating the required conditional probabilities from other empirically derivable probabilities (e.g. Duda and Hart, 1973; Gelman *et al*, 1995). Bayes' theorem for real valued data variables is of the form

$$P\left(C_i|\mathbf{x}\right) = \frac{p\left(\mathbf{x}|C_i\right)P\left(C_i\right)}{p\left(\mathbf{x}\right)} \qquad (6)$$

where $P\left(C_i|\mathbf{x}\right)$, is the *posterior probability*, $p\left(\mathbf{x}|C_i\right)$ is the *likelihood* , $P\left(C_i\right)$, is the *prior probability* of class i occurring and $p(\mathbf{x})$ is the *unconditional density function*. These probabilities are estimated from the data.

For a set of *exclusive* classes, the form of $p(\mathbf{x})$ is given by

$$p(\mathbf{x}) = \sum_{i}^{N} p(\mathbf{x} \cap C_i) = \sum_{i}^{N} p(\mathbf{x}|C_i)P(C_i) \qquad (7)$$

as $\boldsymbol{x}$ belongs to a single class only (Appendix A). Equation (7) ensures that the posterior probabilities sum to unity, i.e.,

$$\sum_{i=1}^{N} P(C_i|\mathbf{x}) = 1 \qquad (8)$$

Equation (7) is a special case of the more general case involving non-exclusive classes given by

$$\begin{aligned} p(\mathbf{x}) &= p(\mathbf{x} \cap U) = \\ p\left(\mathbf{x} \cap \left(\bigcup_{r=1}^{N} C_i\right)\right) &= \sum_{i=1}^{N} p(\mathbf{x} \cap C_i) \\ &\quad - \sum_{i<j}^{N} p(\mathbf{x} \cap C_i \cap C_j) \\ &\quad + \sum_{i<j<k}^{N} p(\mathbf{x} \cap C_i \cap C_j \cap C_k) \\ &\quad \vdots \\ &\quad + (-1)^{N-1} p(\mathbf{x} \cap C_1 \cap C_2 \cap \ldots \cap C_N) \end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^{N} p(\mathbf{x}|C_i)P(C_i) \\ &\quad - \sum_{i<j}^{N} p(\mathbf{x}|C_i \cap C_j)P(C_i \cap C_j) \\ &\quad + \sum_{i<j<k}^{N} p(\mathbf{x}|C_i \cap C_j \cap C_k)P(C_i \cap C_j \cap C_k) \qquad (9) \\ &\quad \vdots \\ &\quad + (-1)^{N+1} p(\mathbf{x}|C_1 \cap C_2 \cap \ldots \cap C_N)P(C_1 \cap C_2 \cap \ldots \cap C_N) \end{aligned}$$

where equation (9) ensures that the probability of the union of the classes conditional upon $\boldsymbol{x}$ is unity, i.e every input is classified.

$$P(U|\mathbf{x}) =$$

$$P\left(\left(\bigcup_{r=1}^{N} C_i\right)|\mathbf{x}\right) = \sum_{i=1}^{N} P(C_i|\mathbf{x})$$

$$- \sum_{i<j}^{N} P(C_i \cap C_j|\mathbf{x})$$

$$+ \sum_{i<j<k}^{N} P(C_i \cap C_j \cap C_k|\mathbf{x}) \qquad (10)$$

$$\vdots$$

$$+ (-1)^{N+1} P(C_1 \cap C_2 \cap \ldots \cap C_N|\mathbf{x})$$

$$= 1$$

where $P(C_i \cap C_j|\mathbf{x}) = \dfrac{p(\mathbf{x}|C_i \cap C_j)P(C_i \cap C_j)}{p(\mathbf{x})}$ etc.

Equation (10) reduces to Equation (8) when $C_i \cap C_j = \phi$, i.e. the classes are exclusive giving rise to the usual definition of Bayes' theorem (e.g. Walpole and Myers, 1989):

Given a partition of the event space, $\{B_1, \ldots . B_N\}$ that is $B_i \cap B_j = \phi, \quad \forall i \neq j$, and a set A such that $A \subseteq \bigcup_{k=1}^{N} B_k$, the conditional probability, $P(B_i|A)$ can be written as

$$P(B_i|A) = \frac{P(B_i|A)}{P(A)} = \frac{P(B_i)P(A|B_i)}{\sum_{j}^{N} P(B_j)P(A|B_j)}$$

Note that the condition that $B_i \cap B_j = \phi, \quad \forall i \neq j$ is required.

10. *Posterior* Knowledge Inclusion For Exclusive Classes

The next four sections examine the inclusion of posterior knowledge for the exclusive, independent and dependent class cases where only members of a single type of class are to be excluded. The more general case is examined in Section 14 where it is shown that the three class types can be decoupled and, thus, treated individually.

For exclusive classes, the situation is shown schematically in Figure 10.

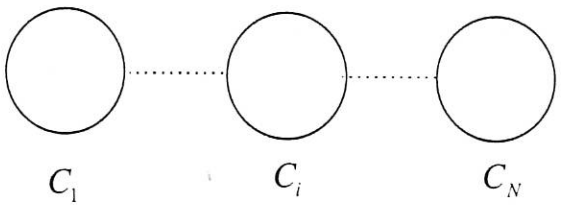

Figure 10. A set of exclusive classes.

Because the classes are exclusive, $P(C_i \cap C_j|\mathbf{x}) = 0 \quad \forall i, j$ i.e. all probabilities of joint classes are zero, only the single class probabilities $P(C_i|\mathbf{x})$ are required to calculate the class union probability, $P\left(\bigcup_{r=1}^{N} C_r|\mathbf{x}\right)$ in Equation (5). This fact leads to the following theorem:

***10.1 Theorem*: Exclusive Class Renormalisation (ECR) Theorem**

For a set of exclusive classes, the updated posterior probabilities, following the exclusion of the set, will be given by a renormalisation of the remaining probabilities.

Proof:

From equation (4)

$$P\left(C_{\delta_i}|\bigcap_k C_{\delta_k}^c \cap \mathbf{x} \cap \varepsilon\right) = \frac{P\left(\bigcup_j C_{\delta_j}|\mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k}|\mathbf{x}\right)}{P\left(\bigcup_l C_{\delta_l}|\mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k}|\mathbf{x}\right)}$$

where $j \in \{\delta_i\} \cup \Delta_\varepsilon$, $k \in \Delta_\varepsilon$, and $l \in \Delta_r \cup \Delta_\varepsilon$.

For the set of exclusive classes, the following equations hold:

$$P\left(\bigcup_j C_{\delta_j}|\mathbf{x}\right) = \sum_j P\left(C_{\delta_j}|\mathbf{x}\right) = P\left(C_{\delta_i}|\mathbf{x}\right) + \sum_k P\left(C_{\delta_k}|\mathbf{x}\right) \qquad (11)$$

$$P\left(\bigcup_k C_{\delta_k}|\mathbf{x}\right) = \sum_k P\left(C_{\delta_k}|\mathbf{x}\right) \qquad (12)$$

and

$$P\left(\bigcup_l C_{\delta_l}|\mathbf{x}\right) = \sum_r P\left(C_{\delta_r}|\mathbf{x}\right) + \sum_k P\left(C_{\delta_k}|\mathbf{x}\right) \qquad (13)$$

where $r \in \Delta_r$

Substituting (11), (12) and (13) into (4) gives

$$P\left(C_{\delta_i} | \bigcap_k C_{\delta_k}^c \cap \mathbf{x} \right) = \frac{P\left(C_{\delta_i} | \mathbf{x}\right)}{\sum_r P\left(C_{\delta_r} | \mathbf{x}\right)} \qquad (14)$$

Equation (14) ensures that

$$\sum_i P\left(C_{\delta_i} | \bigcap_k C_{\delta_k}^c \cap \mathbf{x} \right) = \sum_i \frac{P\left(C_{\delta_i} | \mathbf{x}\right)}{\sum_r P\left(C_{\delta_r} | \mathbf{x}\right)} = \frac{1}{\sum_r P\left(C_{\delta_r} | \mathbf{x}\right)} \sum_i P\left(C_{\delta_i} | \mathbf{x}\right) = 1$$

where $i \in \Delta_r$ ■

10.1.1 Example:

A three class problem is specified as follows:
The input variable $\mathbf{x} = x$ is one dimensional.

Priors: $P(C_1) = P(C_2) = P(C_3) = \frac{1}{3}$

Likelihoods: $P(\mathbf{x}|C_1) \approx N(3,2)$, $P(\mathbf{x}|C_2) \approx N(6,3)$, $P(\mathbf{x}|C_3) \approx N(8,2)$.
Where $N(.,.)$ denotes the normal distribution.

The likelihoods are shown in Figure 11. Note that there are no occurrences of two or more classes together because the classes are exclusive.

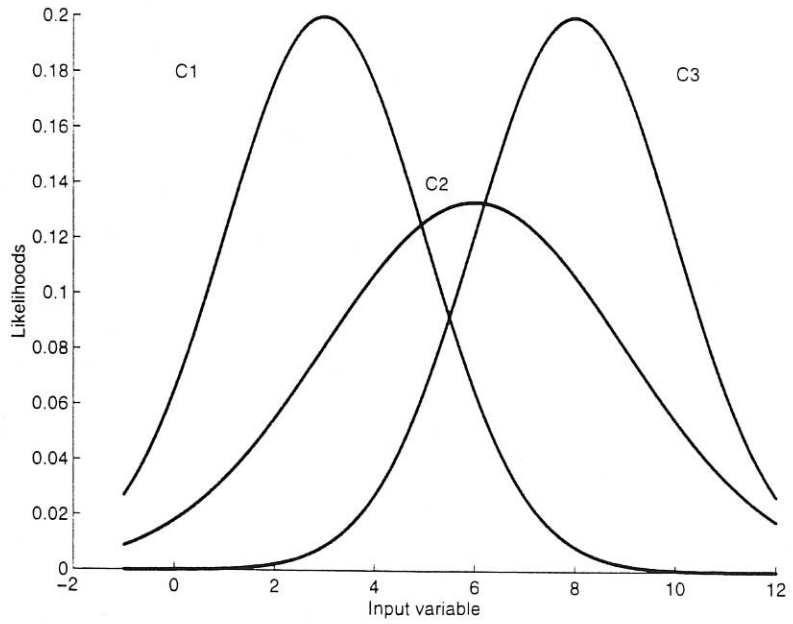

Figure 11 Likelihoods for a three class example where the sets are exclusive.

The posterior probabilities are given by Bayes' theorem

$$P(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)P(C_i)}{p(\mathbf{x})}, \quad i = 1,2,3$$

where

$$p(\mathbf{x}) = \sum_{i}^{3} p(\mathbf{x} \cap C_i) = \sum_{i}^{3} p(\mathbf{x}|C_i)P(C_i)$$

The posterior probabilities are shown in Figure 12.

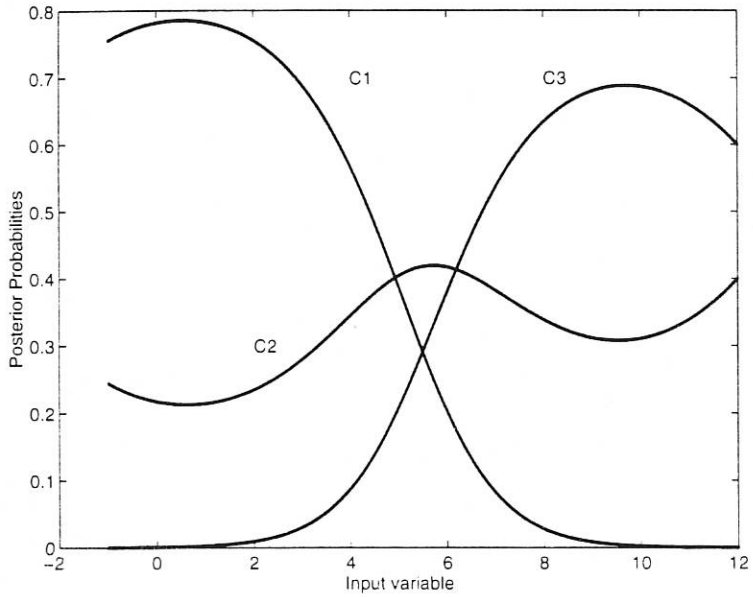

Figure 12.Posterior probabilities for the three class example of Figure 11.

At the point $x = 6$, $P(C_1|6) = 0.2032$, $P(C_2|6) = 0.4172$, $P(C_3|6) = 0.3796$, and

$\sum_{i=1}^{3} P(C_i|\mathbf{x}) = 1$

as expected.

Given the *posterior* knowledge that class three has been excluded, in this case, the updated posterior probabilities are given by equation (4) and shown in Figure 13.

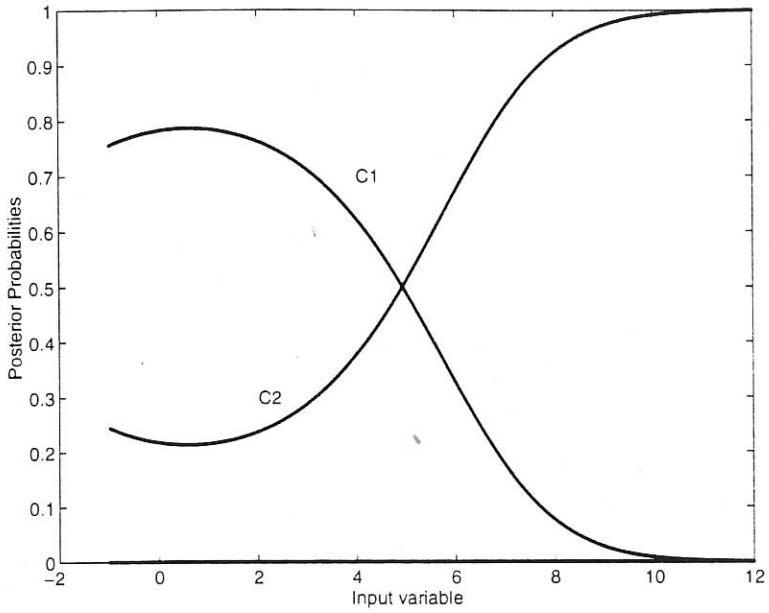

Figure 13. Posterior probabilities where class three has been excluded. At each point, the remaining class probabilities are renormalised.

At the point $x = 6$, $P(C_1|6) = 0.3275$, $P(C_2|6) = 0.6725$.

From the original posterior probabilities, owing to the exclusivness of the classes, by the ECR theorem, there is a simple renormalisation giving

$$P(C_1|6) = \frac{0.2032}{0.2032 + 0.4172} = 0.3275$$

and

$$P(C_1|6) = \frac{0.4172}{0.2032 + 0.4172} = 0.6725$$

as expected.

In Equation (13) it has been assumed that the set of remaining classes (with indices $r \in \Delta_r$) are also exclusive. This is a special case of the more general non-exclusive case. Equation (13) may be modified by substituting $P\left(\bigcup_r C_{\delta_r}|\mathbf{x}\right)$ for $\sum_r P(C_{\delta_r}|\mathbf{x})$ and using Equation (9) to ensure that the probability of the union of the remaining classes is equal to unity.

11. Independence of Classes

So far, the independence of classes has not been dealt with in any detail. This case is also subsumed within equation (4) for updating the posterior probabilities and it will be shown that the exclusion of one or more independent classes does not affect the probability of occurrence of the remaining classes.

What is meant by independence in the context of this paper is *conditional independence* (Bernardo and Smith, 1994; Grimmet and Stirzaker, 1992) (Appendix A.) which involves the independence of posterior probabilities. For example, for pairwise independence we have $P\left(C_i \cap C_j \middle| \mathbf{x}\right) = P\left(C_i \middle| \mathbf{x}\right) P\left(C_j \middle| \mathbf{x}\right)$, and for triplewise independence, $P\left(C_i \cap C_j \cap C_k \middle| \mathbf{x}\right) = P\left(C_i \middle| \mathbf{x}\right) P\left(C_j \middle| \mathbf{x}\right) P\left(C_k \middle| \mathbf{x}\right)$. In general,

$$P\left(\bigcap_l C_l \middle| \mathbf{x}\right) = \prod_l P(C_l \middle| \mathbf{x})$$

The probability of the intersection of the set of classes

$$P\left(\left(\bigcap_s^K C_{\delta_s}\right) \middle| \mathbf{x}\right)$$

can be expanded to give

$$= \frac{P\left(\left(\bigcap_s^K C_{\delta_s}\right) \cap \mathbf{x}\right)}{p(\mathbf{x})}, \text{ by the definition of conditional probability}$$

$$= \frac{P\left(C_{\delta_K} \middle| \left(\bigcap_s^{K-1} C_{\delta_s}\right) \cap \mathbf{x}\right) P\left(\left(\bigcap_s^{K-1} C_{\delta_s}\right) \cap \mathbf{x}\right)}{p(\mathbf{x})}.$$

Continuing this process gives

$$P\left(\left(\bigcap_s^K C_{\delta_s}\right) \middle| \mathbf{x}\right)$$

$$= P\left(C_{\delta_K} \middle| \left(\bigcap_s^{K-1} C_{\delta_s}\right) \cap \mathbf{x}\right) P\left(C_{\delta_{K-1}} \middle| \left(\bigcap_s^{K-2} C_{\delta_s}\right) \cap \mathbf{x}\right) \ldots$$

$$\ldots P\left(C_{\delta_2} \middle| C_{\delta_1} \cap \mathbf{x}\right) P\left(C_{\delta_1} \middle| \mathbf{x}\right) p(\mathbf{x}) \frac{1}{p(\mathbf{x})}$$

$$= P\left(C_{\delta_K} \middle| \left(\bigcap_s^{K-1} C_{\delta_s} \right) \cap \mathbf{x} \right) P\left(C_{\delta_{K-1}} \middle| \left(\bigcap_s^{K-2} C_{\delta_s} \right) \cap \mathbf{x} \right) \ldots$$

$$\ldots P\left(C_{\delta_2} | C_{\delta_1} \cap \mathbf{x} \right) P\left(C_{\delta_1} | \mathbf{x} \right)$$

where there are K members in the set of independent classes.

All set intersections can be decomposed on this way.

Given the revised posterior probability $P\left(C_{\delta_i} | \bigcap_s^K C_{\delta_s}^c \cap \mathbf{x} \right)$ expression of Equation (4), its calculation requires subtracted terms of the form

$P\left(C_{\delta_i} \cap \left(\bigcup_s^K C_{\delta_s} \right) \middle| \mathbf{x} \right) = P\left(\bigcup_s^K \left(C_{\delta_i} \cap C_{\delta_s} \right) \middle| \mathbf{x} \right)$ and

$P\left(\left(\bigcup_r C_{\delta_r} \right) \cap \left(\bigcup_s^K C_{\delta_s} \right) \middle| \mathbf{x} \right) = P\left(\bigcup_s^K \left(\left(\bigcup_r C_{\delta_r} \right) \cap C_{\delta_s} \right) \middle| \mathbf{x} \right)$ which can be decomposed into a sum and difference of terms. For example, take the numerator term

$P\left(\bigcup_s^K \left(C_{\delta_i} \cap C_{\delta_s} \right) \middle| \mathbf{x} \right)$ which can be expanded using terms of the form

$P\left(C_{\delta_i} \cap \left(\bigcap_s^Q C_{\delta_s} \right) \cap \mathbf{x} \right)$ where all terms include C_{δ_i}.

$$P\left(C_{\delta_i} \cap \left(\bigcap_s^Q C_{\delta_s} \right) \cap \mathbf{x} \right) = P\left(C_{\delta_i} \middle| \left(\bigcap_s^Q C_{\delta_s} \right) \cap \mathbf{x} \right) P\left(C_{\delta_Q} \middle| \left(\bigcap_s^{Q-1} C_{\delta_s} \right) \cap \mathbf{x} \right) \ldots$$

$$\ldots P\left(C_{\delta_2} | C_{\delta_1} \cap \mathbf{x} \right) P\left(C_{\delta_1} | \mathbf{x} \right)$$

If *all* classes are independent, for all terms, $P\left(C_{\delta_i} | \bigcap_s^Q C_{\delta_s} \cap \mathbf{x} \right) = P\left(C_{\delta_i} | \mathbf{x} \right)$

$Q \le K$ because $P\left(C_{\delta_i} | \mathbf{x} \right)$ does not depend upon the other classes. Furthermore

$P\left(C_{\delta_Q} \middle| \left(\bigcap_s^{Q-1} C_{\delta_s} \right) \cap \mathbf{x} \right) = P\left(C_{\delta_Q} \middle| \mathbf{x} \right), \forall Q$ giving

$$P\left(C_{\delta_i} \cap \left(\bigcap_s^Q C_{\delta_s} \right) \cap \mathbf{x} \right) = P\left(C_{\delta_i} \middle| \mathbf{x} \right) \prod_s^Q P\left(C_{\delta_i} \middle| \mathbf{x} \right)$$

For this case, all intermediate terms can be calculated from the estimated single-class probabilities e.g. $P\left(C_{\delta_i} \cap C_{\delta_j} | \mathbf{x} \right) = P\left(C_{\delta_i} | \mathbf{x} \right) P\left(C_{\delta_j} | \mathbf{x} \right)$

For I independent classes involved in an intersection,

$$P\left(\left(\bigcap_{s\in\Delta_I} C_{\delta_s}\right)\cap\left(\bigcap_{t\in\Delta_D} C_{\delta_t}\right)\middle|\mathbf{x}\right)$$

$$= P\left(\left(\bigcap_{s\in\Delta_I} C_{\delta_s}\right)\middle|\left(\bigcap_{t\in\Delta_D} C_{\delta_t}\right)\cap\mathbf{x}\right)P\left(\left(\bigcap_{t\in\Delta_D} C_{\delta_t}\right)\middle|\mathbf{x}\right)$$

$$= P\left(\left(\bigcap_{s\in\Delta_I} C_{\delta_s}\right)\middle|\mathbf{x}\right)P\left(\left(\bigcap_{t\in\Delta_D} C_{\delta_t}\right)\middle|\mathbf{x}\right)$$

$$= \left(\prod_{s\in\Delta_I} P\left(C_{\delta_s}|\mathbf{x}\right)\right)P\left(\left(\bigcap_{t\in\Delta_D} C_{\delta_t}\right)\middle|\mathbf{x}\right)$$

For the case of all excluded classes being independent:

11.1 ***Theorem:*** **Independent Class Renormalisation (ICR) Theorem**

For a set of independent classes, the updated posterior probabilities, following the exclusion of this set, will be given by a renormalisation of the remaining probabilities.

Proof:

From equation (4)

$$P\left(C_{\delta_i}|\bigcap_k C^c_{\delta_k}\cap\mathbf{x}\cap\varepsilon\right) = \frac{P\left(\bigcup_j C_{\delta_j}|\mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k}|\mathbf{x}\right)}{P\left(\bigcup_l C_{\delta_l}|\mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k}|\mathbf{x}\right)}$$

where $j\in\{\delta_i\}\cup\Delta_\varepsilon$, $k\in\Delta_\varepsilon$, and $l\in\Delta_r\cup\Delta_\varepsilon$.

Now,

$$P\left(\bigcup_j C_{\delta_j}|\mathbf{x}\right) = P\left(\left(\bigcup_k C_{\delta_k}\right)\cup C_{\delta_i}\middle|\mathbf{x}\right)$$

Further expansion gives

$$P\left(\bigcup_j C_{\delta_j}|\mathbf{x}\right) = P\left(C_{\delta_i}\middle|\mathbf{x}\right) + P\left(\bigcup_k C_{\delta_k}\middle|\mathbf{x}\right) - P\left(\left(\bigcup_k C_{\delta_k}\right)\cap C_{\delta_i}\middle|\mathbf{x}\right) \quad (15)$$

and

$$P\left(\bigcup_l C_{\delta_l} | \mathbf{x}\right) = P\left(\left(\bigcup_k C_{\delta_k}\right) \cap \left(\bigcup_r C_{\delta_r}\right) \middle| \mathbf{x}\right)$$
$$= P\left(\bigcup_r C_{\delta_r} \middle| \mathbf{x}\right) + P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right) - P\left(\left(\bigcup_r C_{\delta_r}\right) \cap \left(\bigcup_k C_{\delta_k}\right) \middle| \mathbf{x}\right) \quad (16)$$

where $r \in \Delta_r$

For independent sets,

$$P\left(\left(\bigcup_k C_{\delta_k}\right) \cup C_{\delta_i} \middle| \mathbf{x}\right) = P\left(C_{\delta_i} | \mathbf{x}\right) P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right) \quad (17)$$

and

$$P\left(\left(\bigcup_r C_{\delta_r}\right) \cap \left(\bigcup_k C_{\delta_k}\right) \middle| \mathbf{x}\right) = P\left(\bigcup_r C_{\delta_r} \middle| \mathbf{x}\right) P\left(\bigcup_k C_{\delta_k} \mathbf{x}\right) \quad (18)$$

substituting (17) and (18) into (15) and (16) respectively gives

$$P\left(\bigcup_j C_{\delta_j} | \mathbf{x}\right) = P\left(C_{\delta_i} | \mathbf{x}\right) + P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right) - P\left(C_{\delta_i} | \mathbf{x}\right) P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right) \quad (19)$$

and

$$P\left(\bigcup_l C_{\delta_l} | \mathbf{x}\right) = P\left(\bigcup_r C_{\delta_r} \middle| \mathbf{x}\right) + P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right) - P\left(\bigcup_r C_{\delta_r} \middle| \mathbf{x}\right) P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right) \quad (20)$$

Finally, substituting (19), and (20) into (4) gives

$$P\left(C_{\delta_i} \middle| \bigcap_k C^c_{\delta_k} \cap \mathbf{x} \cap \varepsilon\right) = \frac{P\left(C_{\delta_i} | \mathbf{x}\right) + P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right) - P\left(C_{\delta_i} | \mathbf{x}\right) P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right)}{P\left(\bigcup_r C_{\delta_r} \middle| \mathbf{x}\right) + P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right) - P\left(\bigcup_r C_{\delta_r} \middle| \mathbf{x}\right) P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right)}$$

giving

$$P\left(C_{\delta_i} \middle| \bigcap_k C^c_{\delta_k} \cap \mathbf{x} \cap \varepsilon\right) = \frac{P\left(C_{\delta_i} | \mathbf{x}\right)}{P\left(\bigcup_r C_{\delta_r} \middle| \mathbf{x}\right)} \quad (21)$$

Where the fact that $P\left(\bigcup_k C_{\delta_k} \middle| \mathbf{x}\right) = 0$ has been used to indicate that these classes have been excluded in this particular case.

Equation (21) ensures that the union of adjusted posterior probabilities is equal to 1.

For the specific case were the remaining sets are exclusive

$$P\left(\bigcup_i C_{\delta_i} \middle| \bigcap_k C^c_{\delta_k} \cap \mathbf{x} \cap \varepsilon\right) = \sum_i P\left(C_{\delta_i} | \bigcap_k C^c_{\delta_k} \cap \mathbf{x} \cap \varepsilon\right) = \sum_i \frac{P\left(C_{\delta_i} | \mathbf{x}\right)}{\sum_r P\left(C_{\delta_r} | \mathbf{x}\right)} = \frac{1}{\sum_r P\left(C_{\delta_r} | \mathbf{x}\right)} \sum_i P\left(C_{\delta_i} | \mathbf{x}\right) = 1$$

where $i \in \Delta_r$ ■

Where excluded classes are independent, the remaining probabilities are renormalised as the excluded classes have no effect on the outcomes.

12. Non-Zero Posterior Probability Adjustment

The independent classes allow *posterior* adjustment of probabilities other than by class exclusion alone. This is apparent by writing the probability for an intermediate term

$$P\left(C_{\delta_i} \cap \left(\bigcap_{s=1}^{Q} C_{\delta_s}\right) \middle| \mathbf{x}\right)$$

$$= P\left(C_{\delta_i} \middle| \left(\bigcap_{s=1}^{Q} C_{\delta_j}\right) \cap \mathbf{x}\right) P\left(\left(\bigcap_{s=1}^{Q} C_{\delta_j}\right) \cap \mathbf{x}\right)$$

$= P\left(C_{\delta_i} | \mathbf{x}\right) P\left(\left(\bigcap_{s=1}^{Q} C_{\delta_s}\right) \cap \mathbf{x}\right)$ where C_{δ_i} is independent of the other classes and $Q \leq K$ where K is the total number of independent classes.

$$P\left(C_{\delta_i} | \mathbf{x}\right) P\left(\left(\bigcap_{s=1}^{Q} C_{\delta_s}\right) \cap \mathbf{x}\right) = P\left(C_{\delta_i} | \mathbf{x}\right) \prod_{s=1}^{Q} P\left(C_{\delta_s} | \mathbf{x}\right)$$

Now, $P\left(C_{\delta_i} | \mathbf{x}\right)$ can be revised (upwards or downwards) and, consequently,

$$P\left(C_{\delta_i} \cap \left(\bigcap_{j=1}^{K} C_{\delta_j}\right) | \mathbf{x}\right).$$

13. *Posterior* Knowledge Inclusion For Non-Independent Classes

For the non-exclusive case, as with the preceding cases involving exclusive and independent classes respectively, only the K conditional probability functions, $P(C_i|\mathbf{x})$ can be evaluated by a learning system as discussed in Sections 18 and 19. Furthermore, even if the K output equations $y_i = f(C_1, C_2, \ldots, C_K, \mathbf{x})$ were independent and contained all the relevant terms (all intersecting class terms), there would be 2^K unknowns in K equations, thus rendering the system of equations indeterminate.

Theorem: The Dependent Class (DC) Theorem

For non-exclusive and dependent classes, neither the ECR Theorem nor the ICR Theorem applies.

Proof;

From (4)

$$P\left(C_{\delta_i}\left|\bigcap_k C^c_{\delta_k} \cap \mathbf{x} \cap \varepsilon\right.\right) = \frac{P\left(\bigcup_j C_{\delta_j}\middle|\mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k}\middle|\mathbf{x}\right)}{P\left(\bigcup_l C_{\delta_l}\middle|\mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k}\middle|\mathbf{x}\right)}$$

where $j \in \{\delta_i\} \cup \Delta_\varepsilon$, $k \in \Delta_\varepsilon$, and $l \in \Delta_r \cup \Delta_\varepsilon$. This expression may be expanded to give

$$P\left(C_{\delta_i}\left|\bigcap_k C^c_{\delta_k} \cap \mathbf{x} \cap \varepsilon\right.\right) = \frac{P\left(C_{\delta_i} \cup \left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k}\middle|\mathbf{x}\right)}{P\left(\left(\bigcup_r C_{\delta_r}\right) \cup \left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k}\middle|\mathbf{x}\right)}$$

$$= \frac{P(C_{\delta_i}|\mathbf{x}) + P\left(\left(\bigcup_k C_{\delta_k}\middle|\mathbf{x}\right)\right) - P\left(C_{\delta_i} \cap \left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k}\middle|\mathbf{x}\right)}{P\left(\left(\bigcup_r C_{\delta_r}\middle|\mathbf{x}\right)\right) + P\left(\left(\bigcup_k C_{\delta_k}\middle|\mathbf{x}\right)\right) - P\left(\left(\bigcup_r C_{\delta_r}\right) \cap \left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k}\middle|\mathbf{x}\right)}$$

$$= \frac{P(C_{\delta_i}|\mathbf{x}) - P\left(C_{\delta_i} \cap \left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right)}{P\left(\left(\bigcup_r C_{\delta_r}\middle|\mathbf{x}\right)\right) - P\left(\left(\bigcup_r C_{\delta_r}\right) \cap \left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right)}$$

$$P\left(C_{\delta_i}\left|\bigcap_k C_{\delta_k}^c \cap \mathbf{x} \cap \varepsilon\right.\right) = \frac{P\left(C_{\delta_i}\left|\mathbf{x}\right.\right) - P\left(C_{\delta_i} \cap \left(\bigcup_k C_{\delta_k}\left|\mathbf{x}\right.\right)\right)}{P\left(\left(\bigcup_r C_{\delta_r}\left|\mathbf{x}\right.\right)\right) - P\left(\left(\bigcup_r C_{\delta_r}\right) \cap \left(\bigcup_k C_{\delta_k}\right)\left|\mathbf{x}\right.\right)}$$

There are intersecting terms in both the numerator and denominator which are non-zero. This precludes using a simple renormalisation to give the revised probabilities. These are only zero for exclusive and independent classes.

Now, for exclusive and independent classes

$$P\left(C_{\delta_i} \cap \left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right) = 0$$

and

$$P\left(\left(\bigcup_r C_{\delta_r}\right) \cap \left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right) = 0$$

which gives

$$P\left(C_{\delta_i}\left|\bigcap_k C_{\delta_k}^c \cap \mathbf{x}\right.\right) = \frac{P\left(C_{\delta_i}\left|\mathbf{x}\right.\right)}{P\left(\left(\bigcup_r C_{\delta_r}\left|\mathbf{x}\right.\right)\right)}$$

as stated by the ECR and ICR formulae.

13.1 An Example:

A three class problem was specified as follows:

Priors:

$$P(C_1) = 0.4,\ P(C_3) = 0.3,\ P(C_2) = 0.1$$

$$P(C_1 \cap C_2) = 0.05$$

$$P(C_2 \cap C_3) = 0.15$$

The likelihoods are given by:

$$P(\mathbf{x}|C_1) = \frac{(0.4N(3.5,2) + 0.05N(4,1))}{0.45}$$

$$P(\mathbf{x}|C_2) = \frac{(0.1N(6,3) + 0.05N(4,1) + 0.15N(5,0.5))}{0.30}$$

$$P(\mathbf{x}|C_3) = \frac{(0.3N(8,2) + 0.15N(5,0.5))}{0.45}$$

$$P(\mathbf{x}|C_1 \cap C_2) = N(4,1)$$

$$P(\mathbf{x}|C_2 \cap C_3) = N(5,0.5)$$

The likelihoods are shown in Figure 14.

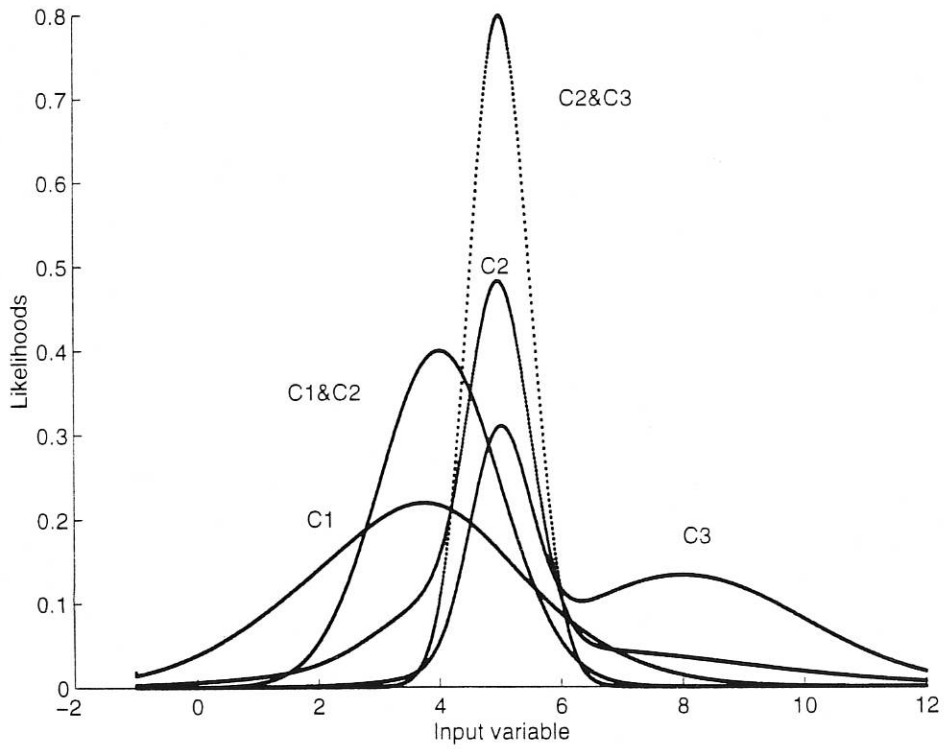

Figure 14. Three class example likelihoods where the classes are not exclusive or independent.

The posterior probabilities are given by Bayes' theorem where

$$P(C_i|\mathbf{x}) = \frac{P(\mathbf{x}|C_i)p(C_i)}{p(\mathbf{x})}, \quad i = 1,2,3$$

and

$$P(C_1 \cap C_2|\mathbf{x}) = \frac{P(\mathbf{x}|C_1 \cap C_2)p(C_1 \cap C_2)}{p(\mathbf{x})}$$

$$P(C_2 \cap C_3|\mathbf{x}) = \frac{P(\mathbf{x}|C_2 \cap C_3)p(C_2 \cap C_3)}{p(\mathbf{x})}$$

and where

$$p(\mathbf{x}) = \sum_i^3 p(\mathbf{x} \cap C_i) - p(\mathbf{x} \cap C_1 \cap C_2) - p(\mathbf{x} \cap C_2 \cap C_3)$$

$$= \sum_i^3 p(\mathbf{x}|C_i)P(C_i) - p(\mathbf{x}|C_1 \cap C_2)P(C_1 \cap C_2) - p(\mathbf{x}|C_2 \cap C_3)P(C_2 \cap C_3)$$

to ensure that

$$p(U|\mathbf{x}) = \sum_i^3 p(C_i|\mathbf{x}) - p(C_1 \cap C_2|\mathbf{x}) - p(C_2 \cap C_3|\mathbf{x}) = 1$$

The posterior probabilities are shown in Figure 15.

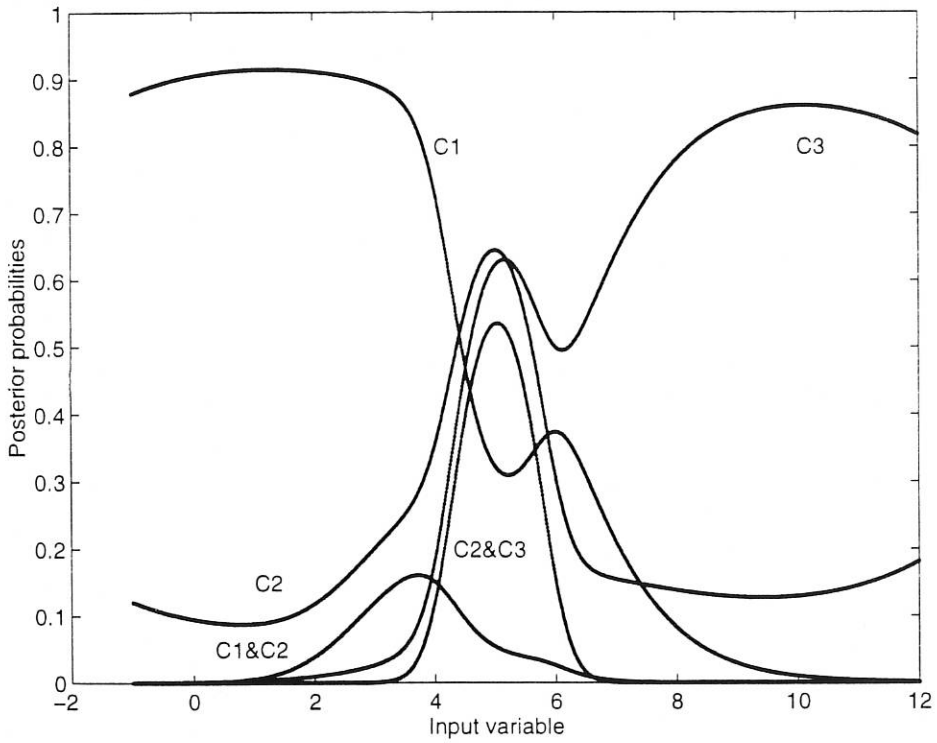

Figure 15. The posterior probabilities for the three class example shown in figure 14.

At the point $x = 5$, $P(C_1|5) = 0.3229$, $P(C_2|5) = 0.6444$, $P(C_3|5) = 0.6210$
and $P(C_1 \cap C_2|5) = 0.0540$, $P(C_2 \cap C_3|5) = 0.5343$
where

$$p(U|5) = \sum_i^3 p(C_i|5) - p(C_1 \cap C_2|5) - p(C_2 \cap C_3|5) = 1$$

as expected.

Given the *posterior* knowledge that $P(C_3|\mathbf{x}) = 0.0$, in this case, the updated posterior probabilities are given by equation (4) and shown in Figure 16.

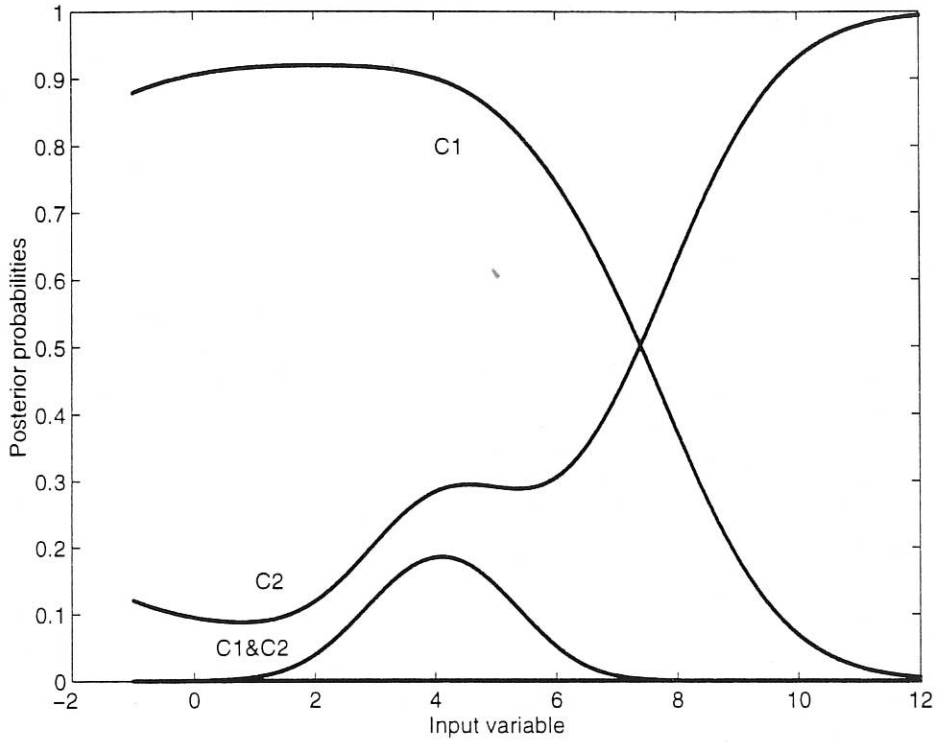

Figure 16. The three class example of figure 14 with class three excluded. This case does not allow a renormalisation as the exclusion of the joint probabilities affect the overall probabilities of the remaining classes.

At the point $x = 5$ $P(C_1|5) = 0.8518$, $P(C_2|5) = 0.2907$ and $P(C_1 \cap C_2|5) = 0.2014$

Because of the non-exclusive classes, the ECR theorem does not apply making

$$P(C_1|5) = \frac{0.3229}{0.3229 + 0.6444 - 0.0540} = 0.3536$$

$$P(C_2|5) = \frac{0.6444}{0.3229 + 0.6444 - 0.0540} = 0.7056$$

and

$$P(C_1 \cap C_2|5) = \frac{0.0540}{0.3229 + 0.6444 - 0.0540} = 0.0591$$

incorrect, using the original probabilities, as expected.

The true values are calculated using equation (4).

$$P(C_1|5) = \frac{0.3229}{0.3229 + 0.6444 - 0.0540 - 0.5343} = 0.852$$

$$P(C_2|5) = \frac{0.6444 - 0.5343}{0.3229 + 0.6444 - 0.0540 - 0.5543} = 0.2905$$

and

$$P(C_1 \cap C_2|6) = \frac{0.0540}{0.3229 + 0.6444 - 0.0540 - 0.5343} = 0.1425$$

Figure 17 illustrates the situation schematically.

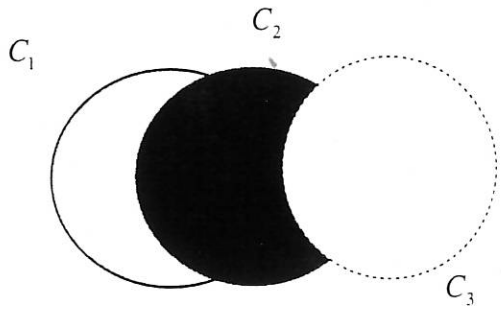

Figure 17 Schematic illustration of a three class problem. A region of class 2 is coupled with a class 3 region; this region will be subtracted from the remaining probabilities for non-independent classes to give the revised posterior probabilities. When class 3 has not occurred, the probability of class 2 is reduced relative to that of class 1.

Table 1 shows the effect of external knowledge on the hierarchy of fault classes.

Prior to External Evidence	Probability	Following External Evidence	Probability
C_2	0.6444	C_1	0.8518
C_3	0.6210	C_2	0.2907
$C_2 \cap C_3$	0.5343	$C_1 \cap C_2$	0.2014
C_1	0.3229	—	
$C_1 \cap C_2$	0.0540	—	

Table 1. The fault class hierarchy both before and after the inclusion of external evidence.

Note that class 1 has risen to the top of the hierarchy following the inclusion of posterior knowledge into the probability adjustment process. A simple renormalisation would have placed class 2 at the top of the hierarchy which would have been incorrect.

14. A Generalised Decoupling of Fault Cases.

So far in sections 10 to 13, the exclusive, independent and dependent cases have been treated separately. The purpose of this section is to deal with the cases taken together and show that the three cases may be decoupled, that is, resolved into separate (non-interacting) sub-processes of the probability update procedure.

Equation (4) gives the revised probabilities:

$$P\left(C_{\delta_i}\left|\bigcap_k C_{\delta_k}^c \cap \mathbf{x} \cap \varepsilon\right.\right) = \frac{P\left(\bigcup_j C_{\delta_j}\left|\mathbf{x}\right.\right) - P\left(\bigcup_k C_{\delta_k}\left|\mathbf{x}\right.\right)}{P\left(\bigcup_l C_{\delta_l}\left|\mathbf{x}\right.\right) - P\left(\bigcup_k C_{\delta_k}\left|\mathbf{x}\right.\right)}$$

The numerator of Equation (4) determines the updated probabilities and is scaled by the denominator. Thus, the numerator makes a convenient starting point for a more general analysis of the three cases, that is, the exclusive, independent and dependent cases taken together.

Recalling that $j \in \{\delta_i\} \cup \Delta_\varepsilon$, $k \in \Delta_\varepsilon$, where Δ_ε is the set of excluded class indices, the numerator is given by:

$$P\left(C_{\delta_i} \cup \left(\bigcup_k C_{\delta_k}\right)\middle|\mathbf{x}\right) - P\left(\bigcup_k C_{\delta_k}\middle|\mathbf{x}\right) \qquad (22)$$

The set of excluded classes can be decomposed into unions of exclusive, independent and dependent classes given by $\bigcup_{r\in\Delta E} C_{\delta_r}$, $\bigcup_{s\in\Delta I} C_{\delta_s}$ and $\bigcup_{t\in\Delta D} C_{\delta_t}$ respectively where ΔE, ΔI and ΔD represent the sets of excluded, independent and dependent class indices respectively. Thus, the set of excluded classes is now given by

$$\bigcup_k C_{\delta_k} = \left(\bigcup_{r\in\Delta E} C_{\delta_r}\right) \cup \left(\bigcup_{s\in\Delta I} C_{\delta_s}\right) \cup \left(\bigcup_{t\in\Delta D} C_{\delta_t}\right) \qquad (23)$$

14.1 Decoupling the Exclusive Fault Classes

Because set exclusivity is the simplest case, the exclusive classes will be dealt with first. This case is straightforward owing to the absence of intersecting classes, i.e. members of the exclusive classes are classified as belonging to a single fault class only. For convenience, equation (23) is represented by

$$\bigcup_k C_{\delta_k} = \left(\bigcup_{r\in\Delta E} C_{\delta_r}\right) \cup \left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right) \qquad (24)$$

where

$$\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)=\left(\bigcup_{s\in\Delta I} C_{\delta_s}\right)\cup\left(\bigcup_{t\in\Delta D} C_{\delta_t}\right) \qquad (25)$$

is the union of non-exclusive independent and dependent classes with indices in the set $\Delta ID = \Delta I \cup \Delta D$.

Equation (22) can now be written as

$$P\left(C_{\delta_i}\cup\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\cup\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)-P\left(\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\cup\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right) \quad (26)$$

Using the set union relation, $P(A\cup B)=P(A)+P(B)-P(A\cap B)$, on both terms of equation (26) gives

$$\left\{P\left(C_{\delta_i}\cup\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\middle|\mathbf{x}\right)+P\left(\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)-P\left(\left[C_{\delta_i}\cup\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\right]\cap\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)\right\}$$
$$-\left\{P\left(\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\middle|\mathbf{x}\right)+P\left(\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)-P\left(\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\cap\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)\right\}$$

which gives a denominator of the form

$$P\left(C_{\delta_i}\cup\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\middle|\mathbf{x}\right)-P\left(\left[C_{\delta_i}\cup\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\right]\cap\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)-P\left(\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\middle|\mathbf{x}\right)$$

where

$$P\left(\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\cap\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)=\phi$$

because of set exclusivity.

Expanding further gives

$$P\left(C_{\delta_i}\middle|\mathbf{x}\right)+P\left(\bigcup_{r\in\Delta E} C_{\delta_r}\middle|\mathbf{x}\right)-P\left(C_{\delta_i}\cap\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\middle|\mathbf{x}\right)$$
$$-P\left(\left[C_{\delta_i}\cup\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\right]\cap\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)-P\left(\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\middle|\mathbf{x}\right)$$

and

$$P\left(C_{\delta_i}\middle|\mathbf{x}\right)-P\left(\left[C_{\delta_i}\cup\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\right]\cap\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right) \qquad (27)$$

where

$$P\left(C_{\delta_i} \cap \left(\bigcup_{r \in \Delta E} C_{\delta_r}\right) \middle| \mathbf{x}\right) = \phi$$

again, because of set exclusivity. Expanding (27) further using the set distributivity property gives

$$P\left(C_{\delta_i} \middle| \mathbf{x}\right) - P\left(\left[C_{\delta_i} \cap \left(\bigcup_{w \in \Delta ID} C_{\delta_w}\right)\right] \cup \left[\left(\bigcup_{r \in \Delta E} C_{\delta_r}\right) \cap \left(\bigcup_{w \in \Delta ID} C_{\delta_w}\right)\right] \middle| \mathbf{x}\right)$$

giving a denominator of

$$P\left(C_{\delta_i} \middle| \mathbf{x}\right) - P\left(\left[C_{\delta_i} \cap \left(\bigcup_{w \in \Delta ID} C_{\delta_w}\right)\right] \middle| \mathbf{x}\right) \qquad (28)$$

By insertion of cancelling terms of the form $P\left(\bigcup_{w \in \Delta ID} C_{\delta_w}\right)$ into Equation (28) it is clear that

$$P\left(C_{\delta_i} \middle| \mathbf{x}\right) + P\left(\bigcup_{w \in \Delta ID} C_{\delta_w}\right) - P\left(\left[C_{\delta_i} \cap \left(\bigcup_{w \in \Delta ID} C_{\delta_w}\right)\right] \middle| \mathbf{x}\right) - P\left(\bigcup_{w \in \Delta ID} C_{\delta_w}\right)$$

gives the form

$$= P\left(\left[C_{\delta_i} \cup \left(\bigcup_{w \in \Delta ID} C_{\delta_w}\right)\right] \middle| \mathbf{x}\right) - P\left(\bigcup_{w \in \Delta ID} C_{\delta_w}\right) \qquad (29)$$

The resulting numerator is for a probability update when there are no exclusive classes to be excluded. The significance of Equation (29) is that a simple renormalisation of the probabilities involving the remaining classes to be excluded on the basis of external knowledge is valid. This is expected for exclusive classes as discussed in Section 10. Here, the more general form indicates that the exclusive class case can be decoupled from the remaining two cases. That is, the separate treatment of exclusive classes as detailed in section 10 is valid.

14.2 Decoupling the Independent Fault Classes

Continuing with the evaluation of the denominator, expanding equation (28) using Equation (25) gives

$$P\left(C_{\delta_i} \middle| \mathbf{x}\right) - P\left(\left(C_{\delta_i} \cap \left(\bigcup_{s \in \Delta I} C_{\delta_s}\right)\right) \cup \left(C_{\delta_i} \cap \left(\bigcup_{t \in \Delta D} C_{\delta_t}\right)\right) \middle| \mathbf{x}\right)$$

$$= P\left(C_{\delta_i} \middle| \mathbf{x}\right) - \left\{P\left(C_{\delta_i} \cap \left(\bigcup_{s \in \Delta I} C_{\delta_s}\right) \middle| \mathbf{x}\right) + P\left(C_{\delta_i} \cap \left(\bigcup_{t \in \Delta D} C_{\delta_t}\right) \middle| \mathbf{x}\right) - P\left(\left(C_{\delta_i} \cap \left(\bigcup_{s \in \Delta I} C_{\delta_s}\right)\right) \cap \left(C_{\delta_i} \cap \left(\bigcup_{t \in \Delta D} C_{\delta_t}\right)\right) \middle| \mathbf{x}\right)\right.$$

giving

$$P\left(C_{\delta_t}\middle|\mathbf{x}\right)-P\left(C_{\delta_t}\cap\left(\bigcup_{s\in\Delta I}C_{\delta_s}\right)\middle|\mathbf{x}\right)-P\left(C_{\delta_t}\cap\left(\bigcup_{t\in\Delta D}C_{\delta_t}\right)\middle|\mathbf{x}\right)+P\left(\left(C_{\delta_t}\cap\left(\bigcup_{s\in\Delta I}C_{\delta_s}\right)\right)\cap\left(C_{\delta_t}\cap\left(\bigcup_{t\in\Delta D}C_{\delta_t}\right)\right)\middle|\mathbf{x}\right)$$

For the second term of this expression,

$$P\left(C_{\delta_t}\cap\left(\bigcup_{s\in\Delta I}C_{\delta_s}\right)\middle|\mathbf{x}\right)$$

$$=\frac{P\left(C_{\delta_t}\cap\left(\bigcup_{s\in\Delta I}C_{\delta_s}\right)\cap\mathbf{x}\right)}{P(\mathbf{x})}$$

by the definition of conditional probability.

Expanding this term further using the chain rule of conditional probability (Krause and Clark, 1993) gives:

$$\frac{P\left(\left(\bigcup_{s\in\Delta I}C_{\delta_s}\right)\middle|C_{\delta_t}\cap\mathbf{x}\right)P\left(C_{\delta_t}\middle|\mathbf{x}\right)P(\mathbf{x})}{P(\mathbf{x})}$$

$$=P\left(\left(\bigcup_{s\in\Delta I}C_{\delta_s}\right)\middle|C_{\delta_t}\cap\mathbf{x}\right)P\left(C_{\delta_t}\middle|\mathbf{x}\right)$$

The set union $\bigcup_{s\in\Delta I}C_{\delta_s}$ is independent of all other sets by definition giving the multiplicative rule

$$P\left(C_{\delta_t}\cap\left(\bigcup_{s\in\Delta I}C_{\delta_s}\right)\middle|\mathbf{x}\right)=P\left(\bigcup_{s\in\Delta I}C_{\delta_s}\middle|\mathbf{x}\right)P\left(C_{\delta_t}\middle|\mathbf{x}\right)$$

Similarly for term 4:

$$P\left(\left(C_{\delta_t}\cap\left(\bigcup_{s\in\Delta I}C_{\delta_s}\right)\right)\cap\left(C_{\delta_t}\cap\left(\bigcup_{t\in\Delta D}C_{\delta_t}\right)\right)\middle|\mathbf{x}\right)$$

$$=P\left(\left(\left(\bigcup_{s\in\Delta I}C_{\delta_s}\right)\cap C_{\delta_t}\cap\left(\bigcup_{t\in\Delta D}C_{\delta_t}\right)\right)\middle|\mathbf{x}\right)$$

$$=P\left(\left(\bigcup_{s\in\Delta I}C_{\delta_s}\right)\middle|C_{\delta_t}\cap\left(\bigcup_{t\in\Delta D}C_{\delta_t}\right)\cap\mathbf{x}\right)P\left(\left(C_{\delta_t}\cap\left(\bigcup_{t\in\Delta D}C_{\delta_t}\right)\right)\middle|\mathbf{x}\right)$$

giving the multiplicative rule

$$P\left(\left(\bigcup_{s\in\Delta I} C_{\delta_s}\right)\cap\left(C_{\delta_i}\cap\left(\bigcup_{t\in\Delta D} C_{\delta_t}\right)\right)\middle|\mathbf{x}\right)=P\left(\bigcup_{s\in\Delta I} C_{\delta_s}\middle|\mathbf{x}\right)P\left(C_{\delta_i}\cap\left(\bigcup_{t\in\Delta D} C_{\delta_t}\right)\middle|\mathbf{x}\right)$$

where the independence property has been used once again as in term 2.

The union of sets $\bigcup_{s\in\Delta I} C_{\delta_s}$ has been excluded on the basis of posterior knowledge

therefore for this single instance given external knowledge, $P\left(\left(\bigcup_{s\in\Delta I} C_{\delta_s}\right)\middle|\mathbf{x}\right)=0$ giving

a final numerator of

$$P\left(C_{\delta_i}\middle|\mathbf{x}\right)-P\left(C_{\delta_i}\cap\left(\bigcup_{t\in\Delta D} C_{\delta_t}\right)\middle|\mathbf{x}\right) \quad (30)$$

The remaining numerator terms in equation (30) indicate that the independent class case has also been decoupled from the exclusive and dependent class cases. The remaining probabilities have been renormalised to give the updated class probabilities following posterior knowledge. Now, only the dependent case remains where a simple renormalisation is not applicable. Equation (30) indicates that intersections of the class probabilities of intersections with the excluded dependent classes must be subtracted to account for dependencies.

For the denominators of Equation (4):

$$P\left(\bigcup_{l} C_{\delta_l}|\mathbf{x}\right)-P\left(\bigcup_{k} C_{\delta_k}|\mathbf{x}\right)$$

Defining the union of remaining sets by $\bigcup_{q\in\Delta R} C_{\delta_q}$, the initial denominator can be written as

$$P\left(\left(\bigcup_{q\in\Delta R} C_{\delta_q}\right)\cup\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\cup\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)-P\left(\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\cup\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)$$

Following a similar analysis to that of the numerator:

$$P\left(\left(\bigcup_{q\in\Delta R} C_{\delta_q}\right)\cup\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\cup\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)-P\left(\left(\bigcup_{r\in\Delta E} C_{\delta_r}\right)\cup\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)$$

$$=P\left(\left(\bigcup_{q\in\Delta R} C_{\delta_q}\right)\cup\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)-P\left(\left(\bigcup_{w\in\Delta ID} C_{\delta_w}\right)\middle|\mathbf{x}\right)$$

which, with Equation (29) gives an updated probability of

$$P\left(C_{\delta_i}\middle|\bigcap_k C_{\delta_k}\cap\mathbf{x}\right)=\frac{P\left(\left[C_{\delta_i}\cup\left(\bigcup_{w\in\Delta ID}C_{\delta_w}\right)\right]\middle|\mathbf{x}\right)-P\left(\bigcup_{w\in\Delta ID}C_{\delta_w}\middle|\mathbf{x}\right)}{P\left(\left[\left(\bigcup_{q\in\Delta R}C_{\delta_q}\right)\cup\left(\bigcup_{w\in\Delta ID}C_{\delta_w}\right)\right]\middle|\mathbf{x}\right)-P\left(\bigcup_{w\in\Delta ID}C_{\delta_w}\middle|\mathbf{x}\right)} \qquad (31)$$

for the set of exclusive classes excluded, Equation (16) is simply a renormalisation following the exclusion of all exclusive classes determined by the posterior knowledge.

Similarly, when the independent classes are excluded,

$$P\left(C_{\delta_i}\middle|\bigcap_k C_{\delta_k}\cap\mathbf{x}\cap\varepsilon\right)=\frac{P\left(\left[C_{\delta_i}\cup\left(\bigcup_{t\in\Delta D}C_{\delta t}\right)\right]\middle|\mathbf{x}\right)-P\left(\bigcup_{t\in\Delta D}C_{\delta_t}\middle|\mathbf{x}\right)}{P\left(\left[\left(\bigcup_{q\in\Delta R}C_{\delta_q}\right)\cup\left(\bigcup_{t\in\Delta D}C_{\delta_t}\right)\right]\middle|\mathbf{x}\right)-P\left(\bigcup_{t\in\Delta D}C_{\delta_t}\middle|\mathbf{x}\right)}$$

where the remaining probabilities have been renormalised once again.

For the remaining case,

$$P\left(C_{\delta_i}\middle|\bigcap_k C_{\delta_k}\cap\mathbf{x}\cap\varepsilon\right)=\frac{P\left(\left[C_{\delta_i}\cup\left(\bigcup_{t\in\Delta D}C_{\delta t}\right)\right]\middle|\mathbf{x}\right)-P\left(\bigcup_{t\in\Delta D}C_{\delta_t}\middle|\mathbf{x}\right)}{P\left(\left[\left(\bigcup_{q\in\Delta R}C_{\delta_q}\right)\cup\left(\bigcup_{t\in\Delta D}C_{\delta_t}\right)\right]\middle|\mathbf{x}\right)-P\left(\bigcup_{t\in\Delta D}C_{\delta_t}\middle|\mathbf{x}\right)}$$

$$=\frac{P\left(C_{\delta_i}\middle|\mathbf{x}\right)-P\left(\left[C_{\delta_i}\cap\left(\bigcup_{t\in\Delta D}C_{\delta t}\right)\right]\middle|\mathbf{x}\right)}{P\left(\left(\bigcup_{q\in\Delta R}C_{\delta_q}\right)\middle|\mathbf{x}\right)-P\left(\left[\left(\bigcup_{q\in\Delta R}C_{\delta_q}\right)\cap\left(\bigcup_{t\in\Delta D}C_{\delta_t}\right)\right]\middle|\mathbf{x}\right)}$$

where $P\left(\left[C_{\delta_i}\cap\left(\bigcup_{t\in\Delta D}C_{\delta t}\right)\right]\middle|\mathbf{x}\right)$ and $P\left(\left[\left(\bigcup_{q\in\Delta R}C_{\delta_q}\right)\cap\left(\bigcup_{t\in\Delta D}C_{\delta_t}\right)\right]\middle|\mathbf{x}\right)$ have to be subtracted from the numerator and denominator respectively because of dependency relations.

14.3 Example: a five set problem:

Two fault classes C_{f1} and C_{f2} whose posterior probabilities will be updated when post-constraint knowledge becomes available. Three classes C_e, C_i and C_d which are exclusive, independent and dependent respectively (that is, $C_e\cap C_i=\phi$ etc.) are to be excluded on the basis of external knowledge.

14.3.1 Excluding the Exclusive Class

The updated probability for class C_{f1} is given by Equation (4)

$$P\left(C_{f1}\middle|C_e^c \cap C_i^c \cap C_d^c \cap \mathbf{x} \cap \varepsilon\right) = \frac{P\left(C_{f1} \cup C_e \cup C_i \cup C_d\middle|\mathbf{x}\right) - P\left(C_e \cup C_i \cup C_d\middle|\mathbf{x}\right)}{P\left(C_{f1} \cup C_{f2} \cup C_e \cup C_i \cup C_d\middle|\mathbf{x}\right) - P\left(C_e \cup C_i \cup C_d\middle|\mathbf{x}\right)}$$

$$= \frac{P\left(\left(C_{f1} \cup C_e\right) \cup \left(C_i \cup C_d\right)\middle|\mathbf{x}\right) - P\left(C_e \cup \left(C_i \cup C_d\right)\middle|\mathbf{x}\right)}{P\left(\left(C_{f1} \cup C_{f2} \cup C_e\right) \cup \left(C_i \cup C_d\right)\middle|\mathbf{x}\right) - P\left(C_e \cup \left(C_i \cup C_d\right)\middle|\mathbf{x}\right)}$$

$$= \frac{P\left(\left(C_{f1} \cup C_e\right) \cup \left(C_{id}\right)\middle|\mathbf{x}\right) - P\left(C_e \cup \left(C_{id}\right)\middle|\mathbf{x}\right)}{P\left(\left(C_{f1} \cup C_{f2} \cup C_e\right) \cup \left(C_{id}\right)\middle|\mathbf{x}\right) - P\left(C_e \cup \left(C_{id}\right)\middle|\mathbf{x}\right)}$$

$$= \frac{P\left(C_{f1} \cup C_e\middle|\mathbf{x}\right) + P\left(C_{id}\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_e\right) \cap C_{id}\middle|\mathbf{x}\right) - \left\{P\left(C_e\middle|\mathbf{x}\right) + P\left(C_{id}\middle|\mathbf{x}\right) - P\left(C_e \cap C_{id}\middle|\mathbf{x}\right)\right\}}{P\left(C_{f1} \cup C_{f2} \cup C_e\middle|\mathbf{x}\right) + P\left(C_{id}\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_{f2} \cup C_e\right) \cap C_{id}\middle|\mathbf{x}\right) - \left\{P\left(C_e\middle|\mathbf{x}\right) + P\left(C_{id}\middle|\mathbf{x}\right) - P\left(C_e \cap C_{id}\middle|\mathbf{x}\right)\right.}$$

giving

$$P\left(C_{f1}\middle|C_e^c \cap C_i^c \cap C_d^c \cap \mathbf{x} \cap \varepsilon\right) = \frac{P\left(C_{f1} \cup C_e\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_e\right) \cap C_{id}\middle|\mathbf{x}\right) - P\left(C_e\middle|\mathbf{x}\right) + P\left(C_e \cap C_{id}\middle|\mathbf{x}\right)}{P\left(\left(C_{f1} \cup C_{f2}\right) \cup C_e\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_{f2} \cup C_e\right) \cap C_{id}\middle|\mathbf{x}\right) - P\left(C_e\middle|\mathbf{x}\right) + P\left(C_e \cap C_{id}\right.}$$

$$= \frac{P\left(C_{f1} \cup C_e\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_e\right) \cap C_{id}\middle|\mathbf{x}\right) - P\left(C_e\middle|\mathbf{x}\right) + P(\phi)}{P\left(\left(C_{f1} \cup C_{f2}\right) \cup C_e\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_{f2} \cup C_e\right) \cap C_{id}\middle|\mathbf{x}\right) - P\left(C_e\middle|\mathbf{x}\right) + P(\phi)}$$

$$= \frac{P\left(C_{f1} \cup C_e\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_e\right) \cap C_{id}\middle|\mathbf{x}\right) - P\left(C_e\middle|\mathbf{x}\right)}{P\left(\left(C_{f1} \cup C_{f2}\right) \cup C_e\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_{f2} \cup C_e\right) \cap C_{id}\middle|\mathbf{x}\right) - P\left(C_e\middle|\mathbf{x}\right)}$$

Expanding the first term of both the numerator and denominator and cancelling terms

$$P\left(C_{f1}\middle|C_e^c \cap C_i^c \cap C_d^c \cap \mathbf{x} \cap \varepsilon\right) = \frac{P\left(C_{f1}\middle|\mathbf{x}\right) - P\left(C_{f1} \cap C_e\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_e\right) \cap C_{id}\middle|\mathbf{x}\right)}{P\left(C_{f1} \cup C_{f2}\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_{f2}\right) \cap C_e\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_{f2} \cup C_e\right) \cap C_{id}\middle|\mathbf{x}\right)}$$

Expanding further and recalling that C_e is exclusive of all other classes gives

$$P\left(C_{f1}\middle|C_e^c \cap C_i^c \cap C_d^c \cap \mathbf{x} \cap \varepsilon\right) = \frac{P\left(C_{f1}\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cap C_{id}\right) \cup \left(C_e \cap C_{id}\right)\middle|\mathbf{x}\right)}{P\left(C_{f1} \cup C_{f2}\middle|\mathbf{x}\right) - P\left(\left(\left(C_{f1} \cup C_{f2}\right) \cap C_{id}\right) \cup \left(C_e \cap C_{id}\right)\middle|\mathbf{x}\right)} \quad (32)$$

$$= \frac{P\left(C_{f1}\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cap C_{id}\right)\middle|\mathbf{x}\right)}{P\left(C_{f1} \cup C_{f2}\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_{f2}\right) \cap C_{id}\middle|\mathbf{x}\right)}$$

which is the updated probability for C_{f1} given the posterior knowledge that classes C_i and C_d have not occurred. The effect of class, C_e has been decoupled from the probability update problem giving,

$$P\left(C_{f1}\middle|C_e^c \cap C_i^c \cap C_d^c \cap \mathbf{x} \cap \varepsilon\right) = \frac{P\left(C_{f1}\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cap C_{id}\right)\middle|\mathbf{x}\right)}{P\left(C_{f1} \cup C_{f2}\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cup C_{f2}\right) \cap C_{id}\middle|\mathbf{x}\right)}$$

which represents the updated posterior probabilities with no exclusive classes.

14.3.2 Excluding the Independent Class

Now the case where the independent class is decoupled from the remaining dependent class will be dealt with. Independence in this case means that C_i is independent of the remaining classes.

Expanding the numerator using $C_{id} = C_i \cup C_d$ gives

$$P\left(C_{f1}\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cap C_i\right) \cup \left(C_{f1} \cap C_d\right)\middle|\mathbf{x}\right)$$

leading to

$$P\left(C_{f1}\middle|\mathbf{x}\right) - \left\{P\left(\left(C_{f1} \cap C_i\right)\middle|\mathbf{x}\right) + P\left(\left(C_{f1} \cap C_d\right)\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cap C_i\right) \cap \left(C_{f1} \cap C_d\right)\middle|\mathbf{x}\right)\right\}$$

Using the chain rule of conditional probability (Krause and Clark, 1993) on terms 2 and 4 gives a numerator of

$$P\left(C_{f1}\middle|\mathbf{x}\right) - P\left(C_i\middle|C_{f1} \cap \mathbf{x}\right)P\left(C_{f1}\middle|\mathbf{x}\right) - P\left(C_{f1} \cap C_d\middle|\mathbf{x}\right) + P\left(C_i\middle|C_{f1} \cap C_d \cap \mathbf{x}\right)P\left(C_{f1} \cap C_d\middle|\mathbf{x}\right)$$

Recalling that class C_i is independent of the remaining classes, and carrying out a similar analysis for the denominator, gives

$$P\left(C_{f1}\middle|C_e^c \cap C_i^c \cap C_d^c \cap \mathbf{x} \cap \varepsilon\right) = \frac{P\left(C_{f1}\middle|\mathbf{x}\right) - P\left(C_{f1}\middle|\mathbf{x}\right)P\left(C_i\middle|\mathbf{x}\right) - P\left(C_{f1} \cap C_d\middle|\mathbf{x}\right) + P\left(C_i\middle|\mathbf{x}\right)P\left(C_{f1} \cap C_d\middle|\mathbf{x}\right)}{P\left(C_{f1} \cup C_{f2}\middle|\mathbf{x}\right) - P\left(C_{f1} \cup C_{f2}\middle|\mathbf{x}\right)P\left(C_i\middle|\mathbf{x}\right) - P\left(\left(\left(C_{f1} \cup C_{f2}\right) \cap C_d\right)\middle|\mathbf{x}\right) + P\left(C_i\middle|\mathbf{x}\right)P\left(\left(C_{f1} \cup C_{f2}\right) \cap C_d\middle|\mathbf{x}\right)}$$

The independent class $P\left(C_i\middle|\mathbf{x}\right)$ has been excluded following posterior knowledge so, setting $P\left(C_i\middle|\mathbf{x}\right) = 0$ to indicate that C_i has been excluded in this particular case gives

$$P\left(C_{f1}\middle|C_e^c \cap C_i^c \cap C_d^c \cap \mathbf{x} \cap \varepsilon\right) =$$

$$\frac{P\left(C_{f1}\middle|\mathbf{x}\right) - P\left(\left(C_{f1} \cap C_d\right)\middle|\mathbf{x}\right)}{P\left(C_{f1} \cup C_{f2}\middle|\mathbf{x}\right) - P\left(\left(\left(C_{f1} \cup C_{f2}\right) \cap C_d\right)\middle|\mathbf{x}\right)}$$

which is the updated probability for C_{f1} given the posterior knowledge that class C_d has not occurred. The effect of class, C_i has also been decoupled from the update problem.

15. Probability Update Procedure

i) Using the estimated priors, determine the *exclusive* fault classes to be excluded on the basis of external knowledge and renormalise the remaining class probabilities,
ii) From the estimated posterior probabilities, determine the *independent* classes to be excluded and renormalise the remaining class probabilities,
iii) finally, use the probability update equation (equation (4)) to exclude the *non-independent* classes.

16. Bounds on the Number of Probability Terms.

Forming the set of all classes $U = \{C_1, \ldots C_N\}$, denoting the number of elements in a set by $|.|$ and denoting the power set of U by $pow(U) = \{\phi, \{C_1\}, \ldots, \{C_N\}, \{C_1, C_2\}, \ldots, \{C_{N-1}, C_N\}, \ldots, \{C_1, \ldots, C_N\}\}$, the number of terms involved in calculating $P\left(\bigcup_{r=1}^{N} C_r\right)$ is now given by $|pow(U)| = 2^N - 1$. This follows, because each member of the power set of U determines uniquely a corresponding probability term in equation 2.

At the worst case, $2^N - 1$ probability distributions must be calculated where N is the number of classes giving complete coverage of all class combinations.

17. An Incremental Version of *Posterior* Knowledge Inclusion

Equation (4) requires that expressions such as $P\left(\left(\bigcup_{s=1}^{K} C_{\delta_s}\right)\middle|\mathbf{x}\right)$ be evaluated in terms of positive probabilities where δ_s signifies the index of a class involved in the union.

When a new item of *posterior* knowledge is made available (i.e. another class is excluded following further inspection) the expression for the updated probabilities, $P\left(C_{\delta_i}\middle|C^c_{\delta_{Nr-1}} \cap C^c_{\delta_{Nr+2}} \cap \ldots \cap C^c_{\delta_N} \cap \mathbf{x} \cap \varepsilon\right)$, becomes

$P\left(C_{\delta_i} | C^c_{\delta_{Nr-1}} \cap C^c_{\delta_{Nr-2}} \cap \ldots \cap C^c_{\delta_N} \cap C_\varepsilon \cap \mathbf{x} \cap \varepsilon\right)$ where C_e is the new excluded class and $e \in \{\delta_1, \ldots, \delta_{Nr}\}$. This requires that the set union is increased by a single set giving $P\left(\bigcup_{s=1}^{K+1} C_{\delta_s} \middle| \mathbf{x}\right)$ for the sets of interest where $\delta_{K+1} = e$.

Using the relation $P(A \cup B|\mathbf{x}) = P(A|\mathbf{x}) + P(B|\mathbf{x}) - P(A \cap B|\mathbf{x})$ and substituting in the expressions

$A = \bigcup_{s=1}^{K} C_{\delta_s}$ and $B = C_e = C_{\delta_{K+1}}$, the updated set union becomes

$$P\left(\left(\bigcup_{s=1}^{K+1} C_{\delta_s}\right)\middle|\mathbf{x}\right) = P\left(\left(\bigcup_{s=1}^{K} C_{\delta_s}\right) \cup C_{\delta_{K+1}} \middle| \mathbf{x}\right)$$
$$= P\left(\left(\bigcup_{s=1}^{K} C_{\delta_s}\right)\middle|\mathbf{x}\right) + P\left(C_{\delta_{K+1}}|\mathbf{x}\right) - P\left(\left(\bigcup_{s=1}^{N} C_{\delta_s}\right) \cap \left(C_{\delta_{K+1}}|\mathbf{x}\right)\right) \quad (33)$$

Expanding using equation (5) gives

$$P\left(\bigcup_{s=1}^{K+1} C_{\delta_s} \middle| \mathbf{x}\right) = \left\{ \begin{array}{l} \sum_{i=1}^{K} P\left(C_{\delta_i}|\mathbf{x}\right) - \sum_{i<j}^{N} P\left(C_{\delta_i} \cap C_{\delta_j}|\mathbf{x}\right) \\ \qquad + \sum_{i<j<k}^{K} P\left(C_{\delta_i} \cap C_{\delta_j} \cap C_{\delta_k}|\mathbf{x}\right) \\ \qquad \vdots \\ \qquad + (-1)^{K+1} P\left(C_{\delta_1} \cap C_{\delta_2} \cap \ldots \cap C_{\delta_K}|\mathbf{x}\right) \end{array} \right\} + P\left(C_{N+1}|\mathbf{x}\right)$$
$$- \left\{ \begin{array}{l} \sum_{i=1}^{K} P\left(C_{\delta_i} \cap C_{\delta_{K+1}}|\mathbf{x}\right) - \sum_{i<j}^{N} P\left(C_{\delta_i} \cap C_{\delta_j} \cap C_{\delta_{K+1}}|\mathbf{x}\right) \\ \qquad + \sum_{i<j<k}^{N} P\left(C_{\delta_i} \cap C_{\delta_j} \cap C_{\delta_k} \cap C_{\delta_{K+1}}|\mathbf{x}\right) \\ \qquad \vdots \\ \qquad + (-1)^{N+1} P\left(C_{\delta_1} \cap C_{\delta_2} \cap \ldots \cap C_{\delta_K} \cap C_{\delta_{K+1}}|\mathbf{x}\right) \end{array} \right\}$$

For incremental inclusion of *posterior* knowledge, the appropriate probability terms must be added or subtracted for each excluded class.

18. The Estimation Problem: 1 from n.

The unprocessed fault diagnosis data will consist of pre-classified input space vectors. The inclusion of *posterior* information requires posterior probabilities to be estimated either directly, or indirectly via Bayes' theorem from this data.

A common method of estimating posterior probabilities is to use an *artificial neural network* (e.g. Bishop, 1995; Richard and Lippmann, 1991). Where the classes are exclusive, given N classes, there arises the 1 from N estimation problem, that is, for each input, one condition class will be chosen on the basis of the posterior probabilities.

The analysis given here is general and applies to both regression and classification problems and involves minimising the mean square error (MSE). For classification problems, however, the cross-entropy measure is more useful and a similar result for cross-entropy will be found in Richard and Lippman (1991).

Assuming discrete outputs indicating class membership, d_{ij} where i signifies the output and j signifies the discrete output value, for L output values, the (MSE) of classification can be calculated by

$$E = \lim_{N_d \to \infty} \frac{1}{N_d} \sum_{n=1}^{N_d} \sum_{i=1}^{N} \sum_{j=1}^{L} \left[y_i - d_{ij} \right]^2 P\left(d_{ij} \cap \mathbf{x} \right) \qquad (34)$$

for N_d data points

For continuous output values and applying the law of large numbers (e.g. Bishop, 1995)

$$E = \sum_{i=1}^{N} \iint \left[y_i - d_i \right]^2 p\left(d_i \cap \mathbf{x} \right) dd_i d\mathbf{x}$$

$$= \sum_{i=1}^{N} \iint \left[y_i - d_i \right]^2 p\left(d_i | \mathbf{x} \right) p(\mathbf{x}) dd_i d\mathbf{x}$$

where d_i is a continuous variable.

From equation (34) using the law of large numbers and assuming that $P\left(d_{ij} \cap \mathbf{x} \right) = P\left(C_j \cap \mathbf{x} \right)$ i.e any network output value depends upon class membership,

$$E = \int \sum_{i=1}^{N} \sum_{j=1}^{N} \left[y_i - d_{ij} \right]^2 p\left(\mathbf{x} \cap C_j \right) d\mathbf{x}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} \int \left[y_i^2 - 2 y_i d_{ij} + d_{ij}^2 \right] p\left(\mathbf{x} \cap C_j \right) d\mathbf{x}$$

$$= \sum_{i=1}^{N} \int \left[y_i^2 \sum_{j=1}^{N} p\left(\mathbf{x} \cap C_j \right) - 2 y_i \sum_{j=1}^{N} d_{ij} p\left(\mathbf{x} \cap C_j \right) + \sum_{j=1}^{N} d_{ij}^2 p\left(\mathbf{x} \cap C_j \right) \right] d\mathbf{x} \qquad (35)$$

For 1 from N classification $d_{ii} = 1$, for $\mathbf{x} \in C_i$ and $d_{ij} = 0$, for $\mathbf{x} \notin C_j$.

Furthermore, as the classes are exclusive (1 from N) and $\boldsymbol{x}$ belongs to one of the classes $\sum_{j=1}^{N} p(\mathbf{x} \cap C_j) = p(\mathbf{x})$, can be substituted into equation (35) to give

$$E = \sum_{i=1}^{N} \int \left[y_i^2 p(\mathbf{x}) - 2 y_i p(\mathbf{x} \cap C_i) + p(\mathbf{x} \cap C_i) \right] d\mathbf{x}$$

$$= \sum_{i=1}^{N} \int \left[y_i^2 p(\mathbf{x}) - 2 y_i P(C_i|\mathbf{x}) p(\mathbf{x}) + P(C_i|\mathbf{x}) p(\mathbf{x}) \right] d\mathbf{x}$$

$$= \sum_{i=1}^{N} \int \left[y_i^2 - 2 y_i P(C_i|\mathbf{x}) + P(C_i|\mathbf{x}) + P^2(C_i|\mathbf{x}) - P^2(C_i|\mathbf{x}) \right] p(\mathbf{x}) d\mathbf{x}$$

$$= \sum_{i=1}^{N} \int \left[y_i^2 - 2 y_i P(C_i|\mathbf{x}) + P(C_i|\mathbf{x}) + P^2(C_i|\mathbf{x}) - P^2(C_i|\mathbf{x}) \right] p(\mathbf{x}) d\mathbf{x}$$

and, finally,

$$E = \sum_{i=1}^{N} \int \left[y_i - P(C_i|\mathbf{x}) \right]^2 p(\mathbf{x}) d\mathbf{x} + \int P(C_i|\mathbf{x}) \left(1 - P(C_i|\mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x}$$

To minimise E with respect to the parameters, the second term can be ignored because it is not a function of the parameters. This leaves the first term which gives $y_i = p(C_i|\mathbf{x})$ for a minimum to occur as the integral will always be positive.

A number of assumptions have been made in the above derivation (Bishop, 1995):

i) a large data set which approximates to an infinite set, is available
ii) parameters (weights) exist such that $y_i(\mathbf{x}, \mathbf{w}) \rightarrow P(C_i|\mathbf{x})$ i.e. the approximating function is able to approximate the required probabilities, and
iii) the optimisation procedure finds the appropriate minimum.

It is also assumed that the classes are exclusive to ensure $\sum_{j=1}^{K} p(\mathbf{x} \cap C_j) = p(\mathbf{x})$. This can be written as

$\sum_{j=1}^{K} P(C_j|\mathbf{x}) p(\mathbf{x}) = p(\mathbf{x})$ which implies that $\sum_{j=1}^{K} P(C_j|\mathbf{x}) = 1$.

19. Conditional Expectation of Vector Output (M from N)

It cannot be assumed that all classes will be exclusive. Where more than one class is likely at any one time, the problem becomes an m from n estimation problem. A procedure analogous to the one above for deriving the result $y_i = p(C_i|\mathbf{x})$ is given below for the more general m from n estimation problem where it is convenient to formulate the class membership problem in terms of vector output.

It will be shown that although joint class information (m from n) is available in the training vectors, a neural network will not be able to estimate the joint probability function unless the output space is expanded to give an equivalent 1 from n problem.

For the continuous valued vector output case (regression problem) the MSE is given by

$$E = \iint \|\mathbf{y} - \mathbf{d}\|^2 p(\mathbf{d} \cap \mathbf{x}) d\mathbf{d} d\mathbf{x} = \iint \|\mathbf{y} - \mathbf{d}\|^2 p(\mathbf{d}|\mathbf{x}) p(\mathbf{x}) d\mathbf{d} d\mathbf{x}$$

which implies that for a minimum MSE,

$$\mathbf{y} = \langle \mathbf{d}|\mathbf{x} \rangle = \int \mathbf{d} p(\mathbf{d}|\mathbf{x}) d\mathbf{d}$$

For the discrete output case (classification) which is of more relevance to the m from n problem, the minimum MSE is given by

$$\mathbf{y} = \langle \mathbf{d}|\mathbf{x} \rangle = \sum_{i=1}^{2^N - 1} \mathbf{d}_i P(\mathbf{d}_i|\mathbf{x}) \text{ where } \mathbf{d}_i \text{ is a binary output vector.}$$

Proof:

$$E = \int \sum_{i=1}^{2^N - 1} \|\mathbf{y} - \mathbf{d}_i\|^2 P(\mathbf{d}_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$= \int \sum_{i=1}^{2^N - 1} \left\| \left(\mathbf{y} - \langle \mathbf{d}|\mathbf{x} \rangle\right) + \left(\langle \mathbf{d}|\mathbf{x} \rangle - \mathbf{d}_i\right) \right\|^2 P(\mathbf{d}_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$= \int \sum_{i=1}^{2^N - 1} \left(\mathbf{y} - \langle \mathbf{d}|\mathbf{x} \rangle\right)^2 P(\mathbf{d}_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$+ 2 \int \sum_{i=1}^{2^N - 1} \left(\mathbf{y} - \langle \mathbf{d}|\mathbf{x} \rangle\right)\left(\langle \mathbf{d}|\mathbf{x} \rangle - \mathbf{d}_i\right) P(\mathbf{d}_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$+ \int \sum_{i=1}^{2^N - 1} \left(\langle \mathbf{d}|\mathbf{x} \rangle - \mathbf{d}_i\right)^2 P(\mathbf{d}_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \qquad (36)$$

For term 1:

$$\int \sum_{i=1}^{2^N-1} \left(\mathbf{y} - \langle \mathbf{d}|\mathbf{x} \rangle\right)^2 P(\mathbf{d}_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$= \int \left(\mathbf{y} - \langle \mathbf{d}|\mathbf{x} \rangle\right)^2 \sum_{i=1}^{2^N-1} P(\mathbf{d}_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$= \int \left(\mathbf{y} - \langle \mathbf{d}|\mathbf{x} \rangle\right)^2 .1. p(\mathbf{x}) d\mathbf{x}$$

$$= \int \left(\mathbf{y} - \langle \mathbf{d}|\mathbf{x} \rangle\right)^2 p(\mathbf{x}) d\mathbf{x}$$

For term 2:

$$2\int \sum_{i=1}^{2^N-1} \left(\mathbf{y} - \langle \mathbf{d}|\mathbf{x} \rangle\right)\left(\langle \mathbf{d}|\mathbf{x} \rangle - \mathbf{d}_i\right) P(\mathbf{d}_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$= 2\int \left(\mathbf{y} - \langle \mathbf{d}|\mathbf{x} \rangle\right) \sum_{i=1}^{2^N-1} \left\{\langle \mathbf{d}|\mathbf{x} \rangle P(\mathbf{d}_i|\mathbf{x}) - \mathbf{d}_i P(\mathbf{d}_i|\mathbf{x})\right\} p(\mathbf{x}) d\mathbf{x}$$

$$= 2\int \left(\mathbf{y} - \langle \mathbf{d}|\mathbf{x} \rangle\right) \left\{\langle \mathbf{d}|\mathbf{x} \rangle \sum_{i=1}^{2^N-1} P(\mathbf{d}_i|\mathbf{x}) - \sum_{i=1}^{2^N-1} \mathbf{d}_i P(\mathbf{d}_i|\mathbf{x})\right\} p(\mathbf{x}) d\mathbf{x}$$

$$= 2\int \left(\mathbf{y} - \langle \mathbf{d}|\mathbf{x} \rangle\right) \left\{\langle \mathbf{d}|\mathbf{x} \rangle .1 - \langle \mathbf{d}|\mathbf{x} \rangle\right\} p(\mathbf{x}) d\mathbf{x}$$

$$= 0$$

For term 3:

$$\int \sum_{i=1}^{2^N-1} \left(\langle \mathbf{d}|\mathbf{x} \rangle - \mathbf{d}_i\right)^2 P(\mathbf{d}_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$= \int \sum_{i=1}^{2^N-1} \left(\langle \mathbf{d}|\mathbf{x} \rangle^2 - 2\langle \mathbf{d}|\mathbf{x} \rangle \mathbf{d}_i + \mathbf{d}_i^2\right) P(\mathbf{d}_i|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$= \int \left\{\langle \mathbf{d}|\mathbf{x} \rangle^2 \sum_{i=1}^{2^N-1} P(\mathbf{d}_i|\mathbf{x}) - 2\langle \mathbf{d}|\mathbf{x} \rangle \sum_{i=1}^{2^N-1} \mathbf{d}_i P(\mathbf{d}_i|\mathbf{x}) + \sum_{i=1}^{2^N-1} \mathbf{d}_i^2 P(\mathbf{d}_i|\mathbf{x})\right\} p(\mathbf{x}) d\mathbf{x}$$

$$= \int \left\{\langle \mathbf{d}|\mathbf{x} \rangle^2 .1 - 2\langle \mathbf{d}|\mathbf{x} \rangle \langle \mathbf{d}|\mathbf{x} \rangle + \langle \mathbf{d}^2|\mathbf{x} \rangle\right\} p(\mathbf{x}) d\mathbf{x}$$

$$= \int \left\{\langle \mathbf{d}^2|\mathbf{x} \rangle - \langle \mathbf{d}|\mathbf{x} \rangle^2\right\} p(\mathbf{x}) d\mathbf{x}$$

Substituting the above terms into equation (36) gives the expression for the MSE

$$E = \int \left(\mathbf{y} - \langle \mathbf{d}|\mathbf{x} \rangle\right)^2 p(\mathbf{x}) d\mathbf{x} + \int \left(\langle \mathbf{d}^2|\mathbf{x} \rangle - \langle \mathbf{d}|\mathbf{x} \rangle^2 \right) p(\mathbf{x}) d\mathbf{x}$$

As the second term is determined by the data, the minimum MSE will be where $\mathbf{y} = \langle \mathbf{d}|\mathbf{x} \rangle$ in the first term

For three classes there will be $2^3 - 1 = 7$ different binary output vectors and the expected conditional output will be given by

$\langle \mathbf{d}|\mathbf{x} \rangle = \sum_{i=1}^{7} \mathbf{d}_i P\left(\left(C_{\gamma 1} \cap \ldots \cap C_{\gamma n} | \mathbf{x} \right)' \right)$ $n \le 3$ where $\gamma_1 \ldots \gamma n \in \Delta_i$ the set of class indices involved for pattern i. The dash denotes the probability of any class or set of classes occurring exclusively i.e. $P\left(\left(C_3|\mathbf{x} \right)' \right)$ does not include $P\left(\left(C_1 \cap C_3|\mathbf{x} \right)' \right)$ et.c.

$$\langle \mathbf{d}|\mathbf{x} \rangle = \mathbf{d}_1 P\left(\left(C_3|\mathbf{x} \right)' \right) + \mathbf{d}_2 P\left(\left(C_2|\mathbf{x} \right) \right) + \mathbf{d}_3 P\left(\left(C_2 \cap C_3|\mathbf{x} \right)' \right) + \mathbf{d}_4 P\left(\left(C_1|\mathbf{x} \right)' \right) + \mathbf{d}_5 P\left(\left(C_1 \cap C_3|\mathbf{x} \right)' \right)$$
$$+ \mathbf{d}_6 P\left(\left(C_1 \cap C_2|\mathbf{x} \right)' \right) + \mathbf{d}_7 P\left(\left(C_1 \cap C_2 \cap C_3|\mathbf{x} \right)' \right)$$

$$\langle \mathbf{d}|\mathbf{x} \rangle = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} P\left(\left(C_3|\mathbf{x} \right)' \right) + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} P\left(\left(C_2|\mathbf{x} \right)' \right) + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} P\left(\left(C_2 \cap C_3|\mathbf{x} \right)' \right) + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} P\left(\left(C_1|\mathbf{x} \right) \right) + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} P\left(\left(C_1 \cap C_3|\mathbf{x} \right)' \right)$$
$$+ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} P\left(\left(C_1 \cap C_2|\mathbf{x} \right)' \right) + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} P\left(\left(C_1 \cap C_2 \cap C_3|\mathbf{x} \right)' \right)$$

This implies that

$$\begin{aligned} y_1 &= P\left(\left(C_1|\mathbf{x} \right)' \right) + P\left(\left(C_1 \cap C_3|\mathbf{x} \right)' \right) + P\left(\left(C_1 \cap C_2|\mathbf{x} \right)' \right) + P\left(\left(C_1 \cap C_2 \cap C_3|\mathbf{x} \right)' \right) \\ &= \left\{ P(C_1|\mathbf{x}) - P(C_1 \cap C_3|\mathbf{x}) - P(C_1 \cap C_2|\mathbf{x}) + P(C_1 \cap C_2 \cap C_3|\mathbf{x}) \right\} \\ &\quad + \left\{ P(C_1 \cap C_3|\mathbf{x}) - P(C_1 \cap C_2 \cap C_3|\mathbf{x}) \right\} \\ &\quad + \left\{ P(C_1 \cap C_2|\mathbf{x}) - P(C_1 \cap C_2 \cap C_3|\mathbf{x}) \right\} \\ &\quad + P(C_1 \cap C_2 \cap C_3|\mathbf{x}) \\ &= P(C_1|\mathbf{x}) \end{aligned}$$

Similarly, $y_2 = P(C_2|\mathbf{x})$ $y_3 = P(C_3|\mathbf{x})$.

In general, for calculating

$$\mathbf{y} = \langle \mathbf{d} | \mathbf{x} \rangle = \sum_{i=1}^{2^N - 1} \mathbf{d}_i P(\mathbf{d}_i | \mathbf{x})$$

for any output, y_i, the class C_i will occur in $\frac{1}{2}2^N = 2^{n-1}$ terms in the summation because the other class intersections form a partition of C_i i.e.

$$P(C_i|\mathbf{x}) = P(C_i|\mathbf{x}) + P\left(\left(C_i \cap C_j\right)'|\mathbf{x}\right) + P\left(\left(C_i \cap C_j \cap C_k\right)'|\mathbf{x}\right) + \ldots$$

$$+ P\left(\left(C_i \cap C_j \cap C_k \cap \ldots \cap C_{N-1} \cap C_N\right)'|\mathbf{x}\right)$$

Figure 18 shows an example of the results obtainable using a multilayer Perceptron to estimate the posterior probabilities of a given a set of distributions. A data set consisting of 1000 data points was used to generate the graph. Note that the classes are not exclusive or independent. The MLP is an instantiation of the estimation problem and is only able to estimate the singleton class posterior probabilities although joint class data is available (i.e. more than one desired output bit may be active at any one time).

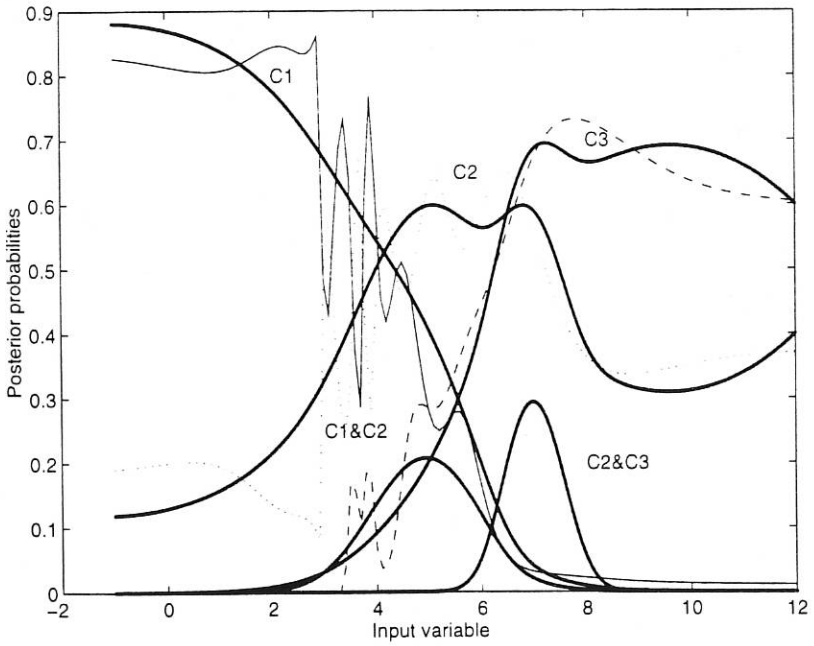

Figure 18 A graph showing the estimation of posterior probabilities by a Multilayer Perceptron. Note that only the singleton class probabilities have been estimated as expected.

It is clear that the probability of class 1 occurring contains some occurrences of class 1 paired with class 2 i.e. $P(C_1 \cap C_2|\mathbf{x}) \neq 0$. Thus, even though the training data

incorporates examples of two classes occurring together, any method using binary outputs to indicate class membership based upon error minimisation as described in sections 18 and 19 (including cross-entropy) is not able to extract this information using N outputs alone where N indicates the number of classes.

To capture class combination information in general, an augmented output vector consisting of $2^N - 1$ outputs is required.

Note the non-smooth approximation of $P(C_1|\mathbf{x})$ and $P(C_2|\mathbf{x})$. This problem and a possible solution, known as regularisation, is discussed in the context of radial basis function networks in section 21.

20. Partitioning the Input Space

The motivation for seeking a partition of the input space is that we need to expand the space to estimate all of the probabilities required for the update equation. In other words, the class dependencies indicated by more than one 'on bit' in the output vector.

A partition of classified input space may be achieved by specifying that the class intersections are pairwise disjoint, for example C_i' only contains data points that belong to C_i and not $C_i \cap C_j$ etc. Similarly, $(C_i \cap C_j)'$ only contains data points that belong to $C_i \cap C_j$ and not $C_i \cap C_j \cap C_k$ etc. This will ensure a partition of the space with disjoint sets as required (e.g. $C_i' \cap (C_i \cap C_j)' = \phi$). The 'dash' notation is used throughout to indicate partition members which compose the entire sample space.

Now, the original formula for the union of sets in terms of set intersections can be specified in purely additive terms:

$$\begin{aligned} P\left(\bigcup_{r=1}^{N} C_r|\mathbf{x}\right) = & \sum_{i=1}^{N} P(C_i'|\mathbf{x}) \\ & + \sum_{\substack{i=1,\, j=2 \\ j\neq k}}^{N} P\left((C_i \cap C_j)'|\mathbf{x}\right) \\ & + \sum_{\substack{i=1,\, j=2,\, k=3 \\ i\neq j\neq k}}^{N} P\left((C_i \cap C_j \cap C_k)'|\mathbf{x}\right) \\ & \vdots \\ & + P\left((C_1 \cap C_2 \cap \ldots \cap C_N)'|\mathbf{x}\right) \end{aligned} \tag{37}$$

It is required to prove that the two representations formally are equivalent.

For $P\left(\bigcup_{r=1}^{N} C_r|\mathbf{x}\right)$, it must be shown that the probability term representing each disjoint region only occurs once in the sum.

For each C_i', $C_i' \subseteq C_i$ and $P(C_i|\mathbf{x})$ occurs only once in the summation $\sum_{i=1}^{N} P(C_i|\mathbf{x})$ and in the expression $P\left(\bigcup_{r=1}^{N} C_r|\mathbf{x}\right)$ because all other class segments consist of two or more intersecting classes and, hence, do not have single class sets as subsets. The first set of terms of $P\left(\bigcup_{r=1}^{N} C_r|\mathbf{x}\right)$ become $\sum_{i=1}^{N} P(C_i'|\mathbf{x})$ where $\sum_{i=1}^{N} P(C_i'|\mathbf{x}) \le \sum_{i=1}^{N} P(C_i|\mathbf{x})$. Introducing the notation $C(n,k)$ which signifies a combination of k objects selected from n. So for C_i' there is only a single set and a single way of selecting that set so n=1 and k=1 giving $C(1,1) = 1$.

For two or more intersecting classes the non-overlapping region of interest is $(C_i \cap C_j)'$. Now, $(C_i \cap C_j)' \subset C_i, C_j, C_i \cap C_j$ so, for

$P(C_i|\mathbf{x}) + P(C_j|\mathbf{x}) - P(C_i \cap C_j|\mathbf{x})$

where all three terms all include the term $P\left(\left((C_i \cap C_j)'|\mathbf{x}\right)\right)$, the resultant term will be $2P\left((C_i \cap C_j)'\middle|\mathbf{x}\right) - P\left((C_i \cap C_j)'\middle|\mathbf{x}\right) = P\left((C_i \cap C_j)'\middle|\mathbf{x}\right)$ i.e. the term only occurs once. Here, the number of terms is given by $C(2,1) - C(2,2) = 1$ where each of the singleton terms $P(C_i|\mathbf{x})$ and $P(C_j|\mathbf{x})$ can be selected once from a set of 2 (because $(C_i \cap C_j)' \subset C_i, C_j$), hence $C(2,1)$, and the term $P\left((C_i \cap C_j)'\middle|\mathbf{x}\right)$ involving 2 sets can only be selected once from $P(C_i \cap C_j|\mathbf{x})$, hence $C(2,2)$. Continuing this argument for $P\left((C_i \cap C_j \cap C_k)'\middle|\mathbf{x}\right)$,

$(C_i \cap C_j \cap C_k)' \subset C_i, C_j, C_k, C_i \cap C_j, C_i \cap C_k, C_j \cap C_k, C_i \cap C_j \cap C_k$ and so the number of terms including $P\left((C_i \cap C_j \cap C_k)'\middle|\mathbf{x}\right)$ will be given by

$C(3,1) - C(3,2) + C(3,3) = 3 - 3 + 1 = 1$. For the general case, n class intersection terms occur $N_n = C(n,1) - C(n,2) + C(n,3) - C(n,4) + \ldots + (-1)^{n+1} C(n,n)$ times. It is required to prove that $N_n = 1$, that is, each term only occurs once.

$$
\begin{aligned}
N_n &= \sum_{k=1}^{n} (-1)^{k+1} C(n,k) \\
&= (-1)\sum_{k=1}^{n} (-1)^{k} C(n,k) \\
&= (-1)\sum_{k=1}^{n} (1)^{n-k}(-1)^{k} C(n,k) \\
&= 1 + (-1) + (-1)\sum_{k=1}^{n} (1)^{n-k}(-1)^{k} C(n,k) \\
&= 1 + (-1)(-1)^{0} + (-1)\sum_{k=1}^{n} (1)^{n-k}(-1)^{k} C(n,k) \\
&= 1 + (-1) + (-1)\sum_{k=0}^{n} (1)^{n-k}(-1)^{k} C(n,k) \\
&= 1 + (1-1)^{n} \\
&= 1
\end{aligned}
$$

Here, the expansion of $(a-b)^n = \sum_{k=1}^{n} (-1)^k a^{n-k} b^k$ has been used with $a = b = 1$. Now any probabilistic function of the possibly overlapping classes $U = \{C_1, \ldots C_N\}$ can be replaced with an equivalent disjoint set $U' = \left\{C_1', \ldots, C_N', (C_1 \cap C_2)', \ldots, (C_{N-1} \cap C_N)', \ldots, (C_1 \cap, \ldots, \cap C_N)'\right\}$ which forms a partition of the input space. Equation (37) can now be written in terms of Bayes theorem:

$$
\begin{aligned}
P\left(\bigcup_{r=1}^{N} C_r | \mathbf{x}\right) = &\sum_{i=1}^{N} \frac{P(C_i')P(\mathbf{x}|C_i')}{P(\mathbf{x})} \\
&+ \sum_{\substack{i=1, j=2 \\ j \neq k}}^{N} \frac{P\left((C_i \cap C_j)'\right) P\left((C_i \cap C_j)' | \mathbf{x}\right)}{P(\mathbf{x})} \\
&+ \sum_{\substack{i=1, j=2, k=3 \\ i \neq j \neq k}}^{N} \frac{P\left((C_i \cap C_j \cap C_k)'\right) P\left(\mathbf{x} | (C_i \cap C_j \cap C_k)'\right)}{P(\mathbf{x})} \\
&\vdots \\
&+ \frac{P\left((C_1 \cap C_2 \cap \ldots \cap C_N)'\right) P\left(\mathbf{x} | (C_1 \cap C_2 \cap \ldots \cap C_N)'\right)}{P(\mathbf{x})}
\end{aligned}
$$

21. Using Radial Basis Function Networks to Estimate the Posterior Probabilities

One way of estimating posterior probabilities is to use a radial basis function network (RBFN) (e.g. Powell, 1987, Broomhead and Lowe, 1988; Moody and Darken, 1989; Bishop, 1995). Radial basis function networks are capable of interpolating between data points to approximate a given noisy function (regression) or probability density function (classification).

A basic RBFN consists of a weighted linear sum of basis functions. This will not be gone into in detail here as there are many references dealing with this subject (e.g. Bishop, 1993, 1995; Haykin, 1994; Wasserman, 1993).

This paper deals with classification problems which necessitates the use of the *softmax function* (e.g Bishop, 1995). To prevent over-learning of the training data, *regularisation* (Bishop, 1991, 1993,1995) may be used. The total cost function for any error-driven neural network using regularisation will be given by

$$C = E + \nu\Omega$$

where E is the original error function, ν is the regularisation constant and Ω is the regularisation function.

For the simulations given below, the second-order differential regularisation function is given by

$$\Omega = \sum_{i=1}^{N}\sum_{l=1}^{L}\left(\frac{\partial^2 y_i}{\partial x_l^2}\right)^2 .$$

Details of the implementation of an RBFN network with second-order differential regularisation applied to a standard network configuration with a softmax layer will be found in Appendix B.

Second-order differential regularisation penalises large changes in the curvature of the output function thus smoothing the resultant function.

The following dependent condition classes were generated using Gaussian distributions for the likelihoods of: C_1, C_2, C_3, $C_1 \cap C_2$, and $C_2 \cap C_3$. The RBFN is expected to approximate the posterior probabilities $P(C_1|\mathbf{x})$, $P(C_2|\mathbf{x})$, $P(C_3|\mathbf{x})$, $P(C_1 \cap C_2|\mathbf{x})$, and $P(C_2 \cap C_3|\mathbf{x})$. The RBFN used had an expanded output set consisting of 5 outputs, each output signifying that case alone e.g. $P(C_1|\mathbf{x})$ gives the posterior probability of class 1 occurring alone. To be consistent with earlier notation: $P(C_i|\mathbf{x}) = P(C_i'|\mathbf{x})$ and $P(C_i \cap C_j|\mathbf{x}) = P\left((C_i \cap C_j)'\middle|\mathbf{x}\right)$.

Figure 19 shows the estimated posterior probabilities without regularisation.

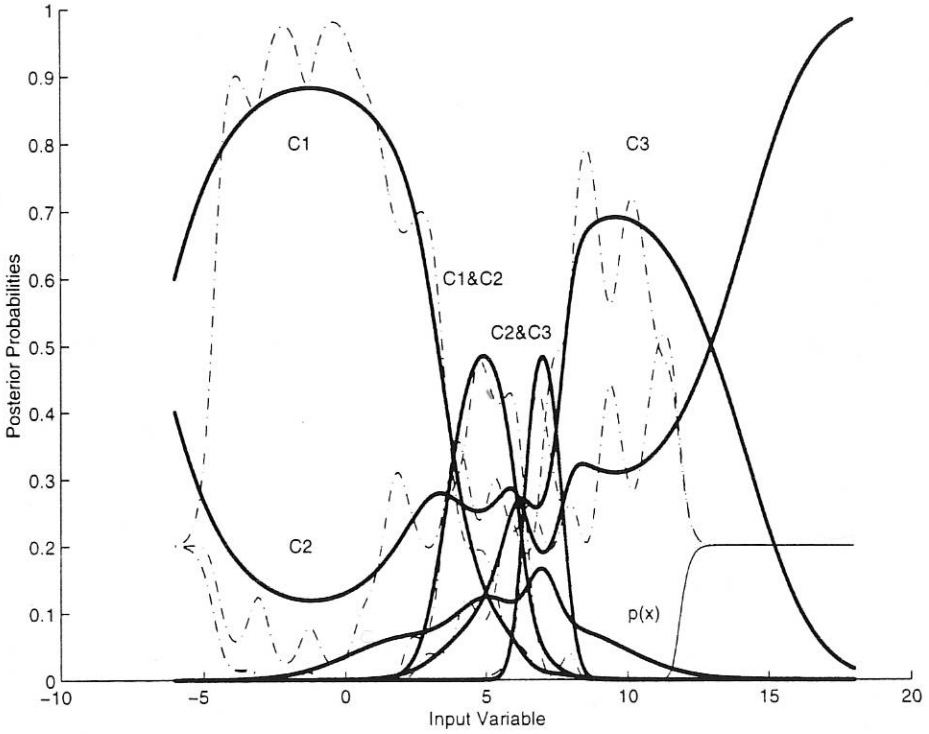

Figure 19. A graph showing the estimation of posterior probabilities by a radial basis function network. Note that only the output space has been partitioned to allow the joint probability functions to be estimated.

The data density outside of the range $[-3,+12]$ is low giving inaccurate predictions of the posterior probability functions as expected. The lack of regularisation allows over-learning of the data and is indicated by the considerable curvature of the estimated probability functions.

Figure 20 shows the estimated posterior probabilities with second-order differential regularisation.

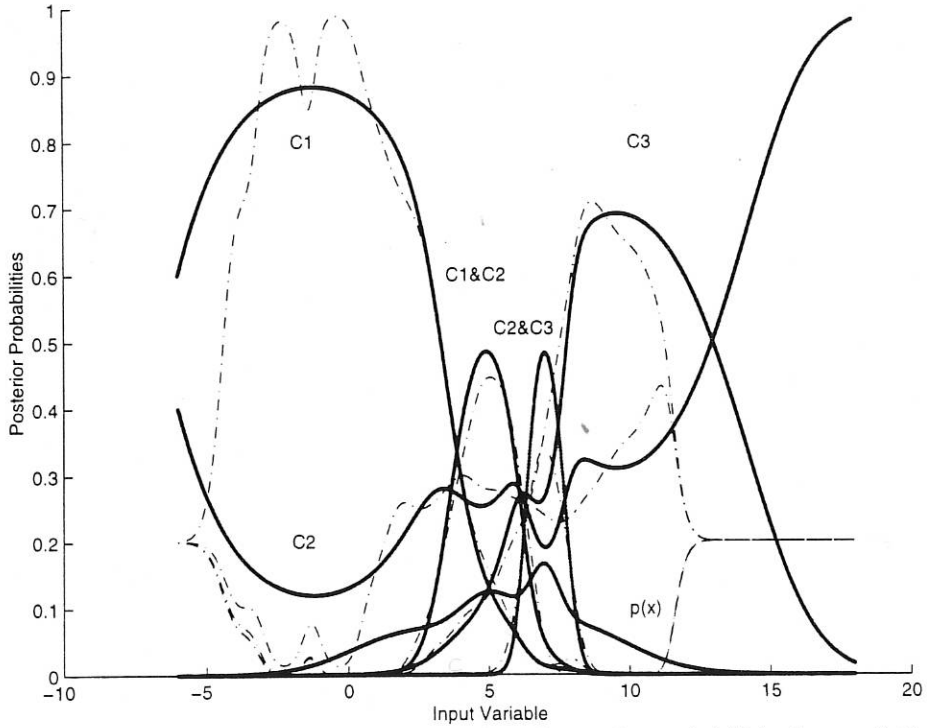

Figure 20. A graph showing the estimation of posterior probabilities by a radial basis function network using second-order differential regularisation as explained in the text.

Note that the approximated functions are considerably smoother in the region of higher data density.

22. Conclusions

Where the posterior knowledge is in the form of evidence indicating the exclusion of classes, posterior probabilities may be revised by a renormalisation of the remaining probabilities for exclusive and independent classes. For non-independent classes, equation (4) may be used.

Posterior knowledge updating requires that a set of posterior probabilities be available, either *a priori* or via estimation. For exclusive and independent classes, only the posterior probabilities of the singleton classes need be known or estimated. For problems in which two or more classes occur simultaneously where the singleton classes are not independent, the joint distributions of posterior probabilities have to be estimated; this entails the use of an augmented output vector to represent the joint probabilities as singleton classes so that they can be estimated. The disadvantage is that the number of outputs (and, thus, probabilities to be estimated) increases exponentially. This could render even modest sized problems difficult to deal with.

However, preprocessing by exclusion of exclusive classes will reduce the complexity to some degree dependent upon the number of exclusive classes. *A priori* knowledge about the independence of some classes will also reduce the problem complexity.

The authors would like to acknowledge the support of both the Engineering and Physical Sciences Research Council of the UK and Rolls-Royce PLC in the production of this work.

References

Applebaum, D (1996) *Probability and Information: An Integrated Approach* CUP Cambridge

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory* John Wiley and Sons Ltd, Chichester.

Bishop, C. M. (1991) Improving the Generalisation Properties of Radial Basis Function Neural Networks, *Neural Computation* **3**, 4, 579-588

Bishop, C. M. (1993) Curvature-Driven Smoothing: A Learning Algorithm For Feedforward Networks *IEEE Transactions on Neural Networks* **4**(5) 882-884

Bishop, C. M.(1995) *Neural networks for Pattern Recognition* Oxford University Press Oxford.

Broomhead, D. S. and Lowe, D. (1988) Multivariable Function Interpolation and Adaptive Networks, *Complex Systems* **2** 321-355

Duda and Hart (1973) *Pattern Classification and Scene Analysis*

Durrett, R. (1994). *The Essentials of Probability* The Duxbury Press. Belmont California.

Haykin, S. (1994) *Neural networks a Comprehensive Foundation* Macmillan

Grimmet, G. R. and Stirzaker, D. R. (1992) *Probability and Random Processes*, Oxford Science Publications, Oxford.

Krause, P. and Clarke, P. (1993) *Representing Uncertain Knowledge: An Artificial Intelligence Approach.* Intellect Books Oxford

Kneale, W. (1949) *Probability and Induction* Oxford at the Clarendon Press, Oxford

Moody, J. and and Darken, C. J. (1989) Fast Learning in Networks of Locally-Tuned Processing Units, *Neural computation* **1** (2) 281-294

Richard, M. D. and Lippmann, R. P. (1991) Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities. *Neural Computation*, **3** 461-483

Walpole, R. E. and Myers, R. H. (1989) *Probability and Statistics for Engineers and Scientists* Macmillan Publishing Company New York

Wasserman, P. D. (1993) *Advanced Methods in Neural Computing* VNR New York

Appendix A

Exclusive sets:

Two classes A and B are *mutually exclusive* or disjoint if $A \cap B = \phi$, that is, if A and B have no elements in common.

Conditional Probability

The *conditional probability* of B, given A, denoted by $P(B|A) = \frac{P(A \cap B)}{P(A)}$ if $P(A) > 0$

Proof of the result

$P(A \cap B^c) = P(A \cup B) - P(B)$

used in the proof of equation (4).

$$
\begin{aligned}
&P(A \cap B^c) = P(A) - P(A \cap B) \\
&= P(A) + P(B) - P(B) - P(A \cap B) \\
&= P(A) + P(B) - P(A \cap B) - P(B) \\
&= P(A \cup B) - P(B)
\end{aligned}
$$

Set Union

This can be proved by induction on n (e.g. Grimmet and Stirzaker, 1992).

$$
\begin{aligned}
P\left(\bigcup_{s=1}^{K} C_s\right) = &\sum_{i=1}^{K} P(C_i) \\
&- \sum_{i<j}^{K} P(C_i \cap C_j) \\
&+ \sum_{i<j<k}^{K} P(C_i \cap C_j \cap C_k) \\
&\vdots \\
&+ (-1)^{K+1} P(C_1 \cap C_2 \cap \ldots \cap C_K)
\end{aligned}
$$

Total Probability

Lemma (Grimmet and Stirzaker, 1992):

For any events A and B

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

More generally, let $B_1, B_2, \ldots, B_N$ be a partition of U. Then,

$$P(A) = \sum_{i=1}^{N} P(A|B_i)P(B_i)$$

Conditional Independence

(Bernardo and Smith, 1994; Grimmet and Stirzaker, 1992)

Definition:

Two events A and B are called *conditionally independent* given C if

$$P(A \cap B|C) = P(A|C)P(B|C).$$

In general, a family of events $\{C_i\}$, $i = 1 \cdots N$ is conditionally independent if

$$P\left(\bigcap_i C_i \middle| \mathbf{x}\right) = \prod_i P(C_i|\mathbf{x})$$

Appendix B.

Radial-Basis Function Network (RBFN) using Cross-Entropy and Second-Order Regularisation

The following analysis is similar to the one carried out for the Multilayer Perceptron in Bishop, 1993.

The Error Function

For a training set of P patterns classified into N classes of conditions, the combined error term consisting of cross-entropy and regularisation components is given by

$$E = \sum_{p=1}^{P} \left\{ E_p^{CE} + \upsilon E_p^R \right\} \qquad \text{(B1)}$$

where the cross-entropy term per pattern is defined as

$$E_p^{CE} = \sum_{n=1}^{N} t_n^p \ln\left(\frac{t_n^p}{y_n^p}\right) \qquad \text{(B2)}$$

and the regularisation term per pattern is given by

$$E_p^R = \frac{1}{2} \sum_{n=1}^{N} \sum_{l=1}^{L} \left(\frac{\partial y_n^p}{\partial \left(x_l^p\right)^2} \right)^2 \qquad \text{(B3)}$$

The RBFN consists of a layer of L input nodes feeding into a layer of J basis function nodes. The layer of J basis function nodes feeds forward into a layer of N output nodes; the N outputs are then fed to softmax function which provides the final outputs.

The final outputs are given by

$$y_i = f\left(a_i\right) \qquad \text{(B4)}$$

where

$$f\left(a_i\right) = \frac{e^{a_i}}{\sum_{k=1}^{N} e^{a_k}} \qquad \text{(B5)}$$

is the softmax function,

$$a_i = \sum_{j=1}^{J} w_{ij} z_j \qquad \text{(B6)}$$

is the net output feeding into the ith output node, and

$$z_j = \phi_j(\mathbf{x}) \qquad \text{(B7)}$$

is the output from the j th basis function.

Gradient descent methods require the calculation of the gradient, $\frac{\partial E}{\partial w_{ij}}$

The gradient can be decomposed to give

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}}\left\{\sum_{p=1}^{P}\left[E_p^{CE} + E_p^R\right]\right\}$$

$$= \sum_{p=1}^{P}\left[\frac{\partial E_p^{CE}}{\partial w_{ij}} + \frac{\partial E_p^R}{\partial w_{ij}}\right]$$

Now, the gradients $\frac{\partial E_p^{CE}}{\partial w_{ij}}$ and $\frac{\partial E_p^R}{\partial w_{ij}}$ defined per pattern are required.

To reduce notational complexity, the superscript p may be dropped.

The Cross-Entropy Gradient Component

Applying the chain rule of differentiation gives

$$\frac{\partial E^{CE}}{\partial w_{ij}} = \frac{\partial E^{CE}}{\partial a_i}\frac{\partial a_i}{\partial w_{ij}} \qquad \text{(B8)}$$

where

$$\frac{\partial E^{CE}}{\partial a_i} = \sum_{i'=1}^{N}\frac{\partial E^{CE}}{\partial y_{i'}}\frac{\partial y_{i'}}{\partial a_i} \qquad \text{(B9)}$$

by applying the chain rule once again.

From Equation (B

$$\frac{\partial E^{CE}}{\partial y_{i'}} = \frac{\partial}{\partial y_{i'}}\left\{\sum_{n=1}^{N} t_n \ln\left(\frac{t_n}{y_n}\right)\right\}$$

$$= t_{i'}.\left(\frac{t_{i'}}{y_{i'}}\right)^{-1}.(-1)\left(y_{i'}\right)^{-2}.t_{i'}$$

giving

$$\frac{\partial E^{CE}}{\partial y_{i'}} = -\frac{t_{i'}}{y_{i'}} \qquad \text{(B10)}$$

$$\frac{\partial y_{i'}}{\partial a_i} = \frac{\partial}{\partial a_i}\left\{\frac{e^{a_{i'}}}{\sum_k^N e^{a_k}}\right\}$$

$$= \frac{\left(\sum_k^N e^{a_k}\right)\frac{\partial}{\partial a_i}e^{a_{i'}} - e^{a_{i'}}e^{a_i}}{\left(\sum_k^N e^{a_k}\right)^2}$$

$$= \frac{e^{a_{i'}}\delta_{ii'}}{\sum_k^N e^{a_k}} - \frac{e^{a_{i'}}}{\sum_k^N e^{a_k}}\frac{e^{a_i}}{\sum_k^N e^{a_k}}$$

giving

$$\frac{\partial y_{i'}}{\partial a_i} = \left(\delta_{ii'} - y_i\right)y_{i'} \qquad \text{(B11)}$$

$$\frac{\partial a_i}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}}\left\{\sum_{j=1}^J w_{ij}z_j\right\} = z_j \qquad \text{(B12)}$$

Substitute equations (B10), (B11) and (B12) into equation (B8)

$$\frac{\partial E^{CE}}{\partial w_{ij}} = \left(\sum_{i'=1}^N \frac{\partial E^{CE}}{\partial y_{i'}}\frac{\partial y_{i'}}{\partial a_i}\right)\frac{\partial a_i}{w_{ij}}$$

$$= \sum_{i'=1}^N \left(-\frac{t_{i'}}{y_{i'}}\right)\left(y_i\delta_{ii'} - y_{i'}y_i\right)z_j$$

$$= \sum_{i'=1}^N \left(-t_i\delta_{ii'} + t_{i'}y_i\right)z_j$$

$$= \left[\left(-\sum_{i'=1}^N t_i\delta_{ii'}\right) + \left(\sum_{i'=1}^N t_{i'}\right)y_i\right]z_j$$

$$= \left(-t_i + 1.y_i\right)z_j$$

giving,

$$\frac{\partial E^{CE}}{\partial w_{ij}} = \left(y_i - t_i\right)z_j \qquad \text{(B13)}$$

The Regularisation Gradient Component

$$\frac{\partial E^R}{w_{ij}} = \sum_{l=1}^{L} \frac{\partial E_l^R}{w_{ij}} \tag{B14}$$

$$\begin{aligned}
\frac{\partial E_l^R}{\partial w_{ij}} &= \sum_n^N \frac{\partial y_n}{\partial x_l^2} \cdot \frac{\partial}{\partial x_l^2}\left(\frac{\partial y_n}{\partial a_{a_i}} z_j\right) \\
&= \sum_n^N \frac{\partial y_n}{\partial x_l^2} \cdot \left[\frac{\partial}{\partial x_l}\left(\frac{\partial}{\partial x_l}\left(\frac{\partial y_n}{\partial a_{a_i}} z_j\right)\right)\right] \\
&= \sum_n^N \frac{\partial y_n}{\partial x_l^2} \cdot \left[\frac{\partial}{\partial x_l}\left(\frac{\partial y_n}{\partial a_{a_i}} \frac{\partial z_j}{\partial x_l} + \frac{\partial}{\partial x_l}\left(\frac{\partial y_n}{\partial a_{a_i}}\right) z_j\right)\right] \\
&= \sum_n^N \frac{\partial y_n}{\partial x_l^2} \left\{\left[\frac{\partial y_n}{\partial a_{a_i}} \frac{\partial}{\partial x_l}\left(\frac{\partial z_j}{\partial x_l}\right) + \frac{\partial z_j}{\partial x_l} \frac{\partial}{\partial x_l}\left(\frac{\partial y_n}{\partial a_{a_i}}\right)\right] + \left[\frac{\partial z_j}{\partial x_l} \frac{\partial}{\partial x_l}\left(\frac{\partial y_n}{\partial a_{a_i}}\right) + \frac{\partial^2}{\partial x_l^2}\left(\frac{\partial y_n}{\partial a_{a_i}}\right) z_j\right]\right\} \\
&= \sum_n^N \frac{\partial y_n}{x_l^2}\left(\frac{\partial y_n}{\partial a_{a_i}}\right)\frac{\partial z_j^2}{\partial x_l^2} + 2\sum_n^N \frac{\partial y_n}{\partial x_l^2} \frac{\partial}{\partial x_l}\left(\frac{\partial y_n}{\partial a_{a_i}}\right)\frac{\partial z_j}{x_l} + \sum_n^N \frac{\partial y_n}{\partial x_l^2} \frac{\partial^2}{\partial x_l^2}\left(\frac{\partial y_n}{\partial a_{a_i}}\right) z_j
\end{aligned}$$

This expression may be rewritten in the form

$$\frac{\partial E_l^R}{\partial w_{ij}} = \sigma_{li} \frac{\partial z_j^2}{\partial x_l^2} + 2\hat{\sigma}_{li} \frac{\partial z_j}{x_l} + \hat{\hat{\sigma}}_{li} z_j \tag{B15}$$

where the following quantities have been defined (Bishop, 1993)

$$\sigma_{li} = \sum_n^N \frac{\partial^2 y_n}{\partial x_l^2}\left(\frac{\partial y_n}{\partial a_{a_i}}\right) \tag{B16}$$

$$\hat{\sigma}_{li} = \sum_n^N \frac{\partial^2 y_n}{\partial x_l^2} \frac{\partial}{\partial x_l}\left(\frac{\partial y_n}{\partial a_{a_i}}\right) \tag{B17}$$

$$\hat{\hat{\sigma}}_{li} = \sum_n^N \frac{\partial^2 y_n}{\partial x_l^2} \frac{\partial^2}{\partial x_l^2}\left(\frac{\partial y_n}{\partial a_{a_i}}\right) \tag{B18}$$

$$\frac{\partial y_n}{\partial a_i} = \left(\delta_{in} - y_i\right) y_n \tag{B19}$$

Now the component derivatives are required in order to evaluate (B15).

The first derivative is

$$\frac{\partial}{\partial x_l}\left(\frac{\partial y_n}{\partial a_i}\right) = \frac{\partial}{\partial x_l}\left(\delta_{in} y_n - y_i y_n\right)$$

$$= \frac{\partial}{\partial x_l}\left(\delta_{in} y_n\right) - \frac{\partial}{\partial x_l}\left(y_i y_n\right)$$

giving

$$\frac{\partial}{\partial x_l}\left(\frac{\partial y_n}{\partial a_i}\right) = \delta_{in}\frac{\partial y_n}{\partial x_l} - y_i\frac{\partial y_n}{\partial x_l} - y_n\frac{\partial y_i}{\partial x_l} \qquad \text{(B20)}$$

which forms a component of (B17). Equation (B20) is differentiated again

$$\frac{\partial^2}{\partial x_l^2}\left(\frac{\partial y_n}{\partial a_i}\right) = \frac{\partial}{\partial x_l}\left(\delta_{in}\frac{\partial y_n}{\partial x_l} - y_i\frac{\partial y_n}{\partial x_l} - y_n\frac{\partial y_i}{\partial x_l}\right)$$

$$= \delta_{in}\frac{\partial^2 y_n}{\partial x_l^2} - y_i\frac{\partial^2 y_n}{\partial x_l^2} - \frac{\partial y_i}{\partial x_l}\frac{\partial y_n}{\partial x_l} - y_n\frac{\partial^2 y_i}{\partial x_l^2} - \frac{\partial y_n}{\partial x_l}\frac{\partial y_i}{\partial x_l}$$

giving

$$\frac{\partial^2}{\partial x_l^2}\left(\frac{\partial y_n}{\partial a_i}\right) = \delta_{in}\frac{\partial^2 y_n}{\partial x_l^2} - y_i\frac{\partial^2 y_n}{\partial x_l^2} - 2\frac{\partial y_i}{\partial x_l}\frac{\partial y_n}{\partial x_l} - y_n\frac{\partial^2 y_i}{\partial x_l^2} \qquad \text{(B21)}$$

which is substituted into (B18).

To evaluate (B20) and (B21) the derivatives $\frac{\partial y_n}{\partial x_l}$ and $\frac{\partial^2 y_n}{\partial x_l^2}$ are required.

Now,

$$\frac{\partial y_n}{\partial x_l} = \sum_{n'=1}^{N}\frac{\partial y_n}{\partial a_{n'}}\frac{\partial a_{n'}}{\partial x_l}$$

$$= \sum_{n'=1}^{N}\left(\delta_{n'n} - y_{n'}\right)y_n\left(\sum_{j'=1}^{J} w_{n'j'}\frac{\partial z_{j'}}{\partial x_l}\right)$$

giving

$$\frac{\partial y_n}{\partial x_l} = \sum_{n'=1}^{N}\sum_{j'=1}^{J} w_{n'j'}\left(\delta_{n'n} - y_{n'}\right)y_n\frac{\partial z_{j'}}{\partial x_l} \qquad \text{(B22)}$$

and,

$$\frac{\partial^2 y_n}{\partial x_l^2} = \frac{\partial}{\partial x_l}\left\{\sum_{n'=1}^{N}\sum_{j'=1}^{J} w_{n'j'}\left(\delta_{n'n} - y_{n'}\right)y_n\frac{\partial z_{j'}}{\partial x_l}\right\}$$

$$= \sum_{n'=1}^{N}\sum_{j'=1}^{J} w_{n'j'}\left\{\frac{\partial}{\partial x_l}\left(\delta_{n'n} - y_{n'}\right).y_n\frac{\partial z_{j'}}{\partial x_l} + \left(\delta_{n'n} - y_{n'}\right)\frac{\partial y_n}{\partial x_l}\frac{\partial z_{j'}}{\partial x_l} + \left(\delta_{n'n} - y_{n'}\right)y_n\frac{\partial^2 z_{j'}}{\partial x_l^2}\right\}$$

giving

$$\frac{\partial^2 y_n}{\partial x_l^2} = \sum_{n'=1}^{N}\sum_{j'=1}^{J} w_{n'j'}\left\{-\frac{\partial y_{n'}}{\partial x_l} y_n \frac{\partial z_{j'}}{\partial x_l} + \left(\delta_{n'n} - y_{n'}\right)\frac{\partial y_n}{\partial x_l}\frac{\partial z_{j'}}{\partial x_l} + \left(\delta_{n'n} - y_{n'}\right) y_n \frac{\partial^2 z_{j'}}{\partial x_l^2}\right\} \quad \text{(B23)}$$

$$\frac{\partial^2 y_n}{\partial x_l^2} = \sum_{n'=1}^{N}\sum_{j'=1}^{J} w_{n'j'}\left\{\left[\left(\delta_{n'n} - y_{n'}\right)\frac{\partial y_n}{\partial x_l} - \frac{\partial y_{n'}}{\partial x_l} y_n\right]\frac{\partial z_{j'}}{\partial x_l} + \left(\delta_{n'n} - y_{n'}\right) y_n \frac{\partial^2 z_{j'}}{\partial x_l^2}\right\}$$

The evaluation of (B22) and (B23) require the derivatives $\frac{\partial z_j}{\partial x_l}$ and $\frac{\partial^2 z_j}{\partial x_l^2}$.

For a specific radial basis function used in this work:

$$z_j = \exp\left(-\frac{\left\|\mathbf{x} - \mathbf{x}_j\right\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\sum_{l=1}^{L}\left(x_l - x_{lj}\right)^2}{2\sigma^2}\right)$$

$$\frac{\partial z_j}{\partial x_l} = -\frac{\left(x_l - x_{lj}\right)}{\sigma^2}\exp\left(-\frac{\sum_{l=1}^{L}\left(x_l - x_{lj}\right)^2}{2\sigma^2}\right) \quad \text{(B24)}$$

$$\frac{\partial^2 z_j}{\partial x_l^2} = \frac{\partial}{\partial x_l}\left(\frac{\partial z_j}{\partial x_l}\right) = \frac{\partial}{\partial x_l}\left\{-\frac{\left(x_l - x_{lj}\right)}{\sigma^2}\exp\left(-\frac{\sum_{l=1}^{L}\left(x_l - x_{lj}\right)^2}{2\sigma^2}\right)\right\}$$

$$= -\frac{\left(x_l - x_{lj}\right)}{\sigma^2}\left[-\frac{\left(x_l - x_{lj}\right)}{\sigma^2}\exp\left(-\frac{\sum_{l=1}^{L}\left(x_l - x_{lj}\right)^2}{2\sigma^2}\right)\right] + \left[-\frac{1}{\sigma^2}\right]\left(\exp\left(-\frac{\sum_{l=1}^{L}\left(x_l - x_{lj}\right)^2}{2\sigma^2}\right)\right)$$

giving

$$\frac{\partial^2 z_j}{\partial x_l^2} = \left[\frac{\left(x_l - x_{lj}\right)^2}{\sigma^4} - \frac{1}{\sigma^2}\right]\exp\left(-\frac{\sum_{l=1}^{L}\left(x_l - x_{lj}\right)^2}{2\sigma^2}\right)$$

The learning rule used to update the network weights was chosen to be a simple gradient descent rule of the form

$$w_{ij}(t+1) = w_{ij}(t) - \eta\frac{\partial E}{\partial w_{ij}}$$

where η is the learning rate. This rule was found to be sufficient to learn the desired probability distributions.