Neural Comput & Applic (1998)7:147–165

Neural
Computing
& Applications

A Decision Support Tool for the Diagnosis of Breast Cancer Based upon Fuzzy ARTMAP

J. Downs[1], R.F. Harrison[1] and S.S. Cross[2]

[1]Department of Automatic Control and Systems Engineering, University of Sheffield, [2]Department of Pathology, University of Sheffield Medical School, Sheffield, UK

This paper presents research into the application of the fuzzy ARTMAP neural network model to the diagnosis of cancer from fine-needle aspirates of the breast. Trained fuzzy ARTMAP networks are differently pruned so as to maximise accuracy, sensitivity and specificity. The differently pruned networks are then employed in a 'cascade' of networks intended to separate cases into 'certain' and 'suspicious' classes. This mimics the predictive behaviour of a human pathologist. The fuzzy ARTMAP model also provides symbolic rule extraction facilities and the validity of the derived rules for this domain is discussed. Additionally, results are provided showing the effects upon network performance of different input features and different observers. The implications of the findings are discussed.

Keywords: Artificial neural networks; Breast cancer; Fuzzy ARTMAP; Symbolic rule extraction

1. Introduction

Neural networks potentially have great value in medical decision-support applications. Unlike expert systems, they bypass the difficult and time-consuming knowledge acquisition process [1] by learning complex associations directly from domain examples. This provides the opportunity for a neural network decision-support tool to adapt to perform the same task under varying conditions. This occurs, for example, because of differing demographic conditions or clinical procedures from region to region, or because procedures may vary over time owing to advances in medical knowledge or technology.

A large and ever-growing body of work now exists on applying neural networks to various medical classification tasks, e.g. the diagnosis of epilepsy, diagnosis of low back disorders, early diagnosis of myocardial infarction, classification of thyroid disorders and identification of Alzheimer's diseased tissue [2–6]. For a general introduction to artificial neural network applications in medicine, see elsewhere [7–9].

The main thrust of this work has been in the use of feedforward networks to learn the association between evidence and outcome. Primarily, the Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) network classes have been employed [10,11]. Both the MLP and the RBF have been shown to be rich enough in structure so as to be able to approximate any (sufficiently smooth) function with arbitrary accuracy [12,13]. In addition, it can be shown that, for one-from-many classification problems, attainment of the minimal value of a variety of cost functions with respect to the weights yields an estimate of the posterior (class conditional) probabilities required for the implementation of a Bayesian classifier [14]. Thus, given sufficient data, computational resources (the MLP, in particular, does not scale well with problem size) and time (non-linear optimisation which is non-linear in the parameters may be time-consuming to perform, numerically), it is possible to estimate the Bayes-optimal classifier to any desired degree of accuracy, directly and with no prior assumptions on the probabilistic structure of the data. However, despite this attractive property, there are two drawbacks with

Correspondence and offprint requests to: R.F. Harrison, Department of Automatic Control and Systems Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK. Email: r.f.harrison@sheffield.ac.uk

these classes of feedforward networks in addition to the caveats given above.

First, these networks require artificial termination of training, since they are susceptible to new but irrelevant data over-writing useful existing associations, and thus degrading general classification performance. However, this requirement compromises the adaptability of a neural network. New data are not always irrelevant, sometimes it reflects significant changes in the classification domain which requires new associations to be learned. This is called the stability-plasticity dilemma: 'How can a learning system be designed to remain plastic, or adaptive, in response to significant events, and yet remain stable in response to irrelevant events?' [15].

The MLP and RBF networks do not cope well with this dilemma. The termination of learning once a pre-determined level of performance has been achieved sacrifices plasticity for the sake of stability. In non-stationary classification domains (i.e. when the underlying statistics of the population are changing with time), these networks cannot incrementally acquire new associations as the environment changes. Instead, they must be completely retrained on new domain data, losing all previously learned associations even though some may still be useful (and will be reacquired alongside the new associations with retraining). Furthermore, when retraining with additional data there is no guarantee that the previous network's topology, learning parameters, etc. will still provide a good solution. It is possible that significant changes to the network will be needed when it is re-derived. (For a detailed discussion on this issue with regard to feedforward networks, see Sharkey and Sharkey [16].)

Many medical domains are non-stationary to a greater or lesser extent, for example, owing to changes in clinical procedures. Furthermore, the artificial termination of learning means that a neural network trained on data from one site is likely to perform the same task sub-optimally using data from another site because of variations in local conditions. Thus, it would be desirable if such a network could be 'fine-tuned' to its changed operating conditions by incremental learning of cases from the new site. In general, causality dictates that it is not possible to know *a priori* whether or not a domain is stationary.

The second problem stems from a common general criticism of the neural network paradigm that the rules governing the predicted outcome are obscure. This can lead to a strong resistance to acceptance of a network's predictions by potential users. This is particularly true for medical domains. For example, a diagnosing clinician using a neural network decision-support tool has to be convinced that the underlying model captures the salient features of the domain and that the system is further able to offer an explanation of its diagnoses in user-comprehensible (i.e. symbolic) terms. However, attempts to extract domain rules from feedforward networks have met with limited success, with, so far, no completely general method published (see elsewhere [17] for a review).

In this paper, we describe the application of a powerful, but relatively little-used, neural network model, fuzzy ARTMAP, to a medical decision-support task – assisting a pathologist in the diagnosis of breast cancer. Fuzzy ARTMAP is a neural network model, using both feedforward and feedback, which is not susceptible to the two criticisms cited above and has other desirable properties for this task (and medical classification tasks in general, see Downs et al. [18]). This is not to say that fuzzy ARTMAP is a panacea, it too has shortcomings, some of which are investigated here. We concentrate here on the ability of the neural network to divulge the set of rules it abstracts during the training process.

Fuzzy ARTMAP, while allowing real valued input data, is not required for binary valued inputs. Thus, the entirely binary system known as ARTMAP [19] would have sufficed here (except when we come to include patient age, as featured in Section 6.1). However, our objective is to investigate the applicability of the system in a real domain and therefore the more general network architecture is discussed and used. It is worth noting that there is a slight computational advantage in implementing the fuzzy system even for binary domains because only one set of weights has to be maintained rather than the two required for ARTMAP.

The structure of the remainder of this paper is as follows. Section 2 introduces the diagnostic problem and considers some previous work in this domain. Section 3 describes the fuzzy ARTMAP neural network model, avoiding the detailed mathematical approach which is amply presented elsewhere, and instead attempting to explain how and why the network works. It then goes on to present associated activities such as voting strategies for overcoming certain problems of the architecture, and finally the extraction of production rules. Section 4 provides basic performance results using fuzzy ARTMAP. Section 5 evaluates the rules derived from the trained networks and Section 6 provides further performance results investigating the effect of variations in input features. Section 7 provides an overall discussion, some conclusions and suggestions for further work.

2. Background to the Diagnostic Problem

2.1. Breast Cancer

Breast cancer is a common disease affecting approximately 22,000 women yearly in England and Wales, and is the commonest cause of death in the 35–55 year age group of the same population [20]. The primary method of diagnosis is through microscopic examination by a pathologist of cytology slides derived from fine needle aspiration of breast lesions (FNAB) [21]. The acquisition of the necessary diagnostic expertise for this task is a relatively slow process. (A trainee pathologist in the UK requires at least five years study and experience before being allowed to sit the final professional pathology examinations for membership of the Royal College of Pathologists.)

Large studies of the cytopathologic diagnosis of FNAB have shown a range of specificity (probability of correctly diagnosing a benign tumour or lesion) of diagnosis of 90%–100% with a range of sensitivities (probability of correctly diagnosing a malign lesion) from 84%–97% [22]. These studies have been produced in centres specialising in the diagnosis of breast disease by pathologists with a special interest in breast cytopathology. In less specialised centres, such as district general hospitals, when a diagnostic FNAB service is being set up, the performance is in the lower range of those values with a specificity of 95% and a sensitivity of 87% [23]. There is thus scope for an artificial intelligence decision-making tool for this domain to assist in training junior pathologists and to improve the performance of experienced pathologists.

The most important performance metric in this domain is not overall diagnostic accuracy but specificity. This is because the pathologist's prime concern is to avoid false positive predictions (diagnosing benign lesions as malignant) since these may result in unnecessary surgery such as mastectomy or wide local excision of the lesion. False negatives are tolerated because, if the clinical suspicion of malignancy remains, the surgeon will then take further samples for additional testing by the pathologist. (Indeed, false negatives are inevitable within this domain since some aspirations fail to locate a malignant lesion and extract nearby healthy tissue.)

In the cytodiagnosis of FNAB there are some observable features which are cited as being important in the recognition of malignant cells. A 'canonical' list is provided by Wells et al. [24], although this publication does not attribute weights to these features or indicate the significance of combinations of these features. Ten of the features described by Wells et al. are utilised in this research. The medical definitions are shown in the appendix, together with the abbreviations by which they will be referred to throughout this paper. As a general guideline, the features NAKED, FOAMY and APOCRINE are regarded as indicators of a benign outcome and all other features are indicative of malignancy. (However, some interactions between conjunctions of features are possible, see Section 5.2.)

2.2. Previous Work on Decision Support for this Domain

Some expert systems have been described which attempt to use human observations of features in FNAB and then apply computers to process these observations and attach weight to the presence and combination of features. Heathfield et al. [25] describe a rule-based expert system with rules derived from cytopathological textbooks and discussions with pathologists but they do not give any results for the performance of the system on a test set of data. A Bayesian belief network has been developed by Hamilton et al. [26]. The conditional probability matrices relating each observed feature to the diagnosis were defined by a cytopathologist. The network was tested using 40 cases, it is difficult to assess the results because four categories of diagnosis were used (benign, malignant, atypical, probably benign and suspicious), but 6% of the true benign cases and 9% of the true malignant cases were assigned to an equivocal category. Wolberg and Mangasarian [22] have produced a large study with a 420 case training set and 215 case test set, and they have used a user-modified computer-generated decision tree, the multi-surface method of pattern separation and a connectionist system with a back-propagation learning algorithm. Nine cytological features were observed and given a scalar value of 1–10. On the test data set the decision tree method gave a specificity of 97% with a sensitivity of 93%, the connectionist network a specificity of 99% and a sensitivity of 97%, the multi-surface separation method produced 100% specificity and sensitivity. However, some cases (such as cancer judged to have been missed by the aspirating needle) were excluded before analysis.

In previous work by the authors [27, 28], the

ARTMAP neural network model was applied to this task using a 313 item training set and a 100 item test set. Various configurations of the model gave an accuracy of 94%–95%, a sensitivity of 90%–96%, and specificity of 92%–99% (for full details, see the earlier work [27]). Atypical cases were not removed prior to measuring system performance. The model was shown to perform at least as well as an expert human pathologist and displayed diagnostic accuracy very close to the optimum possible for the domain. In this paper, we present further results using a new ARTMAP configuration with a revised and expanded data set. The present work also corrects some minor flaws in the methodology of our previous papers.

3. Adaptive Resonance Theory

Adaptive resonance theory, or ART [29] represents a family of neural network models originally developed from the competitive learning paradigm with the intention of overcoming the stability-plasticity dilemma [30]. This was achieved by utilising feedback between layers of input and category nodes in addition to the standard feedforward connections of competitive learning. Thus, in ART models, an input pattern is not automatically assigned to the category that is initially maximally activated by that input. Instead, if the feedback process rejects the initial categorisation, a search process is initiated which terminates when a category node with an acceptable match to the input is found. If no such node exists, a new category node is formed to classify the input.

It should also be noted that ART models usually employ a localist representation for category nodes owing to the so-called 'winner-take-all' competitive learning dynamics. Although biologically implausible, this feature does have the advantage of facilitating symbolic rule extraction from a trained network (see Section 3.5). Furthermore, localisation results from a simplification used to obtain the computational models and is not inherent in adaptive resonance theory *per se* (e.g. see Carpenter [31]).

Since ART was an outgrowth of competitive learning, initial models developed from it employed unsupervised learning. Examples of such models include ART 1 [29] which is restricted to the classification of binary input patterns, and fuzzy ART [32] which generalises ART 1 so as to classify both analogue and binary patterns. More recently, ART models employing supervised learning have been developed which are based upon these earlier models and so retain their self-organising properties.

Fuzzy ARTMAP [33] is one such model, based upon fuzzy ART. It is thus a self-organising, supervised learning, neural network model for the classification of both analogue and binary patterns. Fuzzy ARTMAP consists of three modules, two fuzzy ART systems called ART_a and ART_b, and a related structure called the map field. During training, input patterns are presented to ART_a together with their associated teaching stimuli at ART_b. Associations between patterns at ART_a and ART_b are then formed at the map field. During testing, supervisory inputs at ART_b are omitted, and instead the inputs at ART_a are used to recall a previously learned association with an ART_b pattern via the map field.

However, fuzzy ARTMAP does not directly associate inputs at ART_a and ART_b. Rather, such patterns are first self-organised into prototypical category clusters before being associated at the map field. Hence, generalised associations are formed. If the ART_a category cluster selected through self-organisation does not match with the teaching category at ART_b, the map field generates a re-set at ART_a, forcing the input to be re-classified to an appropriate ART_a category prototype. If no such prototype exists, a new cluster is automatically created for classification of the input. Thus, it can be seen that supervision of learning is only employed when self-organisation leads to a classification error.

Sections 3.1 and 3.2 give a fuller description of the operation of the unsupervised ART network and the supervised fuzzy ARTMAP system, respectively. For those readers whose primary interest lies in the application of the technology, these subsections can be skipped without loss.

3.1. *ART* Operation

The major components of the ART 1 model are depicted in Fig. 1. These components can be grouped into two subsystems: the attentional and orienting subsystems. The F_1 and F_2 field represent layers of

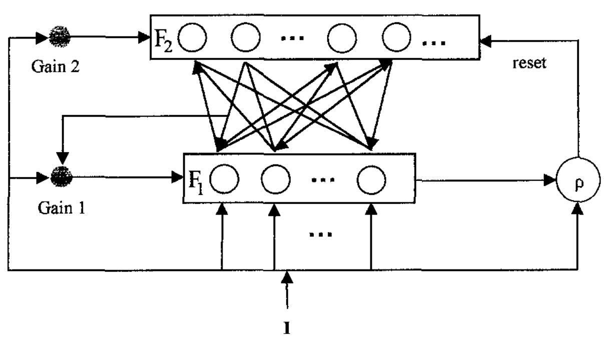

Fig. 1. Schematic of generic ART architecture.

nodes in the attentional subsystem, we will refer them as the *matching* and *category* layers, respectively. Each node in F_1 is connected to every node of F_2 through a set of bottom-up weights. Similarly, each node in F_2 is connected to nodes in F_1 through top–down weights. In addition, the set of nodes comprising the F_2 field are completely connected. The nodes in F_1 and F_2 fields are used to encode patterns of Short-Term Memory (STM) activity, while the weighted connections between the nodes in the F_1 and F_2 fields are used to store Long-Term Memory (LTM) traces.

The orienting subsystem receives input from the input and the F_1 field. It will generate a reset signal to the F_2 field whenever the input pattern is not matched closely enough to the pattern of STM activity across the F_1 field.

When the F_1 field is activated by an input pattern, **I**, a pattern of activation is generated across the nodes of F_1. This pattern represents a pattern in STM. This STM activity across the F_1 field generates an output vector from the field F_1 and sends an inhibitory signal to the orienting subsystem. The network is designed such that this inhibitory signal exactly cancels the excitatory effect of the signal from **I**, so that the orienting subsystem remains inactive and only the attentional portion of the ART network is active. The output vector is multiplied by the bottom–up LTM traces (matrix of adaptive weights), and the result is a bottom–up input vector, which is supplied to the F_2 field. Next, a competition cycle occurs among the F_2 field nodes, so that only the node receiving the maximal input remains active (winner-take-all), generating an STM activity vector across the F_2 field. This STM activity across the F_2 field generates an output vector which is multiplied by the top–down LTM traces to generate a top–down input to the F_1 field. This pattern is called a *learned expectation* or *top–down template*.

Now, both the bottom–up input pattern **I** and the top–down pattern perturb the F_1 field. The nodes in F_1 (and F_2 as well) are designed so that they can become active only if two out of three sources of input are active. This feature is called the 2/3 rule. Thus, only those nodes receiving signals from both **I** and F_2 remain active. At this point, the F_1 field tries to match the top–down signal against **I**, generating a new STM activity vector. If there is a significant mismatch between bottom–up and top–down inputs at the F_1 field, this new STM activity results in a new output activity from the F_1 field which causes attenuation in the total size of the inhibitory signal from F_1 to the orienting subsystem. As a result, if this attenuation is sufficiently great, the orienting subsystem is engaged and a reset signal is released which inhibits the STM activity at the F_2 field. This inhibition leads to the elimination of the top–down input terminating the mismatch. Hence, the initial STM activity is reinstated across the F_1 field. Once again, this STM activity across the F_1 field generates an output from the F_1 field which produces the same bottom–up input at the F_2 field as before. Since the node initially chosen in the F_2 field remains inhibited, a new node in the F_2 field is chosen. If once more the top–down input significantly mismatches the bottom–up at the F_1 field, then a search for an appropriate F_2 field node is continued until a node is found that matches the input pattern, **I**, to the degree of accuracy required by the level of the attentional vigilance, ρ, or a previously uncommitted F_2 node is found, or the system capacity is exhausted and cannot accommodate **I**.

If a reset signal is not generated by the orienting subsystem, then the network settles down into a resonance state of STM activation. It is then that the LTM traces learn any information about the input pattern because the resonant state persists long enough for learning to occur; hence the term *Adaptive Resonance Theory*. For full details of ART operation see elsewhere [29,32].

3.2. Fuzzy *ARTMAP* Operation

Fuzzy ARTMAP makes use of the operations of fuzzy set theory instead of the classical set theory that governs the dynamics of ARTMAP. A FAM network consists of two fuzzy ART modules, ART_a and ART_b, connected by a map field as shown in Fig. 2. During supervised learning, an input pattern vector $\boldsymbol{a}$ is fed to ART_a with its target vector $\boldsymbol{b}$ to ART_b. ART_a and ART_b cluster their input vectors independently. An intervening map field (F_{ab}) adaptively associates predictive antecedents in ART_a with their consequents in ART_b.

In ART_a, an input vector $\boldsymbol{a}$ first registers itself at the F_{1a} layer in complement-coded format, *i.e.* $\boldsymbol{A} = (\boldsymbol{a}, 1\text{-}\boldsymbol{a})$, to avoid the category proliferation problem. This pattern vector is fanned-out to all the nodes in the F_{2a} layer via a set of long term memory weights. The response of each F_{2a} node is based upon a fuzzy *choice function*

$$\frac{|\boldsymbol{A} \wedge \boldsymbol{w}_{a\text{-}j}|}{\alpha_a + |\boldsymbol{w}_{a\text{-}j}|}$$

where $\boldsymbol{w}_{a\text{-}j}$ is the weight vector of the jth F_{2a} node, α_a is the choice parameter of ART_a and the fuzzy 'and' operator ($\wedge$) and the norm $|.|$ are defined as: $(\boldsymbol{x} \wedge \boldsymbol{y})_i = \min(x_i, y_i)$ and $|\boldsymbol{x}| = \sum_i |x_i|$ [9].

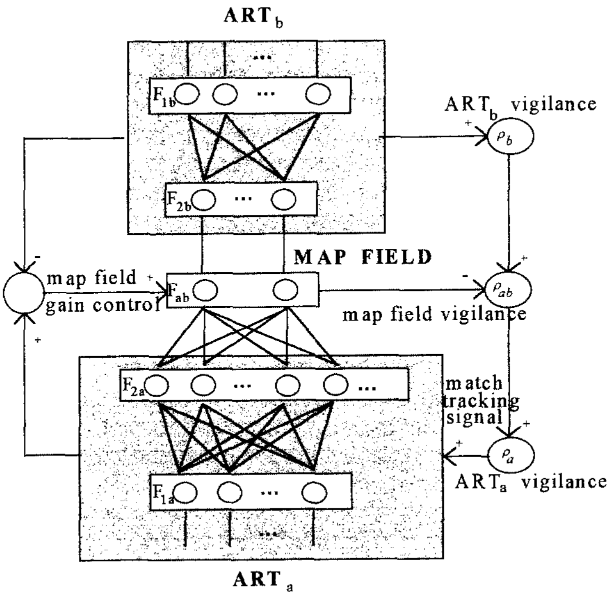

Fig. 2. A schematic diagram of the fuzzy ARTMAP network.

The maximally activated node is selected as the winner and all other nodes are suppressed in accordance with the winner-take-all competitive structure. The winning F_{2a} node then feeds back its weight vector to F_{1a}. This weight vector represents the category prototype of the winning node and is used for comparison with the input vector against a vigilance threshold. Resonance is said to occur if the vigilance test is satisfied, i.e.

$$\frac{|A \wedge w_{a\text{-}J}|}{|A|} \geq \rho_a$$

where ρ_a is the vigilance parameter of ART$_a$ and $w_{a\text{-}J}$ is the winning, Jth, node in F_{2a}. Otherwise, a mismatch signal is sent to F_{2a} to reset the winning node for the rest of the pattern matching cycle. The input vector A is now re-transmitted to F_{2a} to select a new winner. This search cycle ends when the current category prototype is able to meet the vigilance test or a new node is recruited in F_{2a} with the input pattern coded as the prototypical weight vector. An entirely parallel set of operations takes place in ART$_b$, with obvious re-labelling of vectors.

After resonance has occurred in ART$_a$ and ART$_b$, a predictive signal is sent from the winning F_{2a} node to the map field. If this prediction is disconfirmed by the winning node in F_{2b}, i.e. the map field vigilance test fails, a control strategy called match-tracking is initiated. Match-tracking increases ρ_a to a value which triggers a search in ART$_a$. Thus, ρ_a is made slightly greater than $|A \wedge w_{a\text{-}j}| / |A|$ to cause the ART$_a$ vigilance test to fail. In such a way, match-tracking provides a means to select a node in F_{1a} which fulfils both the ART$_a$ and the map field vigilance tests. If such a node does not exist, F_{2a} is *shut down* for the rest of the input presentation.

Additionally, input patterns are initially 'pre-processed' by complement coding, such that each input feature also has its complement passed to the fuzzy ARTMAP input layer. (The fuzzy ARTMAP input layer is thus twice the length of the original input vector). Complement coding is necessary to help reduce proliferation of category clusters in ART models. With purely binary data, complement coding ensures that the vector received by the fuzzy ARTMAP input layer always has a fixed number of true bits (equal to the length of the original input vector).

Training in fuzzy ARTMAP almost always results in multiple category clusters forming at ART$_a$, for each teaching category present at ART$_b$, with each such cluster encoding multiple input exemplars (i.e. each ART$_a$ cluster represents a significant sub-region of the overall state space covered by a particular teaching category). Hence, fuzzy ARTMAP instantiates a many-to-one mapping between ART$_a$ input patterns and their actual classification. For full details of fuzzy ARTMAP, see Carpenter et al. [33].

3.3. Simplified Fuzzy *ARTMAP*

Simplified fuzzy ARTMAP (henceforth abbreviated to SFAM) is a 'streamlined' version of fuzzy ARTMAP intended to be more computationally efficient than a full implementation but with a minimal loss of computational power [34]. Figure 3 gives a diagrammatic representation of the model; circled lines denote adaptive weight connections, arrowed lines show processing flow. The teaching stimulus has a dashed arrow to indicate its variable status – if it is present learning occurs, if it is absent prediction takes place instead.

The model does not self-organise teaching inputs at ART$_b$, but instead encodes these patterns directly. (Thus, unlike fuzzy ARTMAP, the ART$_b$ module in SFAM is not a complete fuzzy ART system.) This is based on the observation that in most pattern classification tasks the teaching stimuli themselves do not need to be further categorised since they directly represent distinct, known classes, e.g. one-from-many classification.

In addition, SFAM converts all but one of the three user-changeable parameters in fuzzy ARTMAP to constants whose values are the usual default settings of the original parameters. (For the benefit of those familiar with the ARTMAP models, the category choice parameter, a, is fixed to be near-

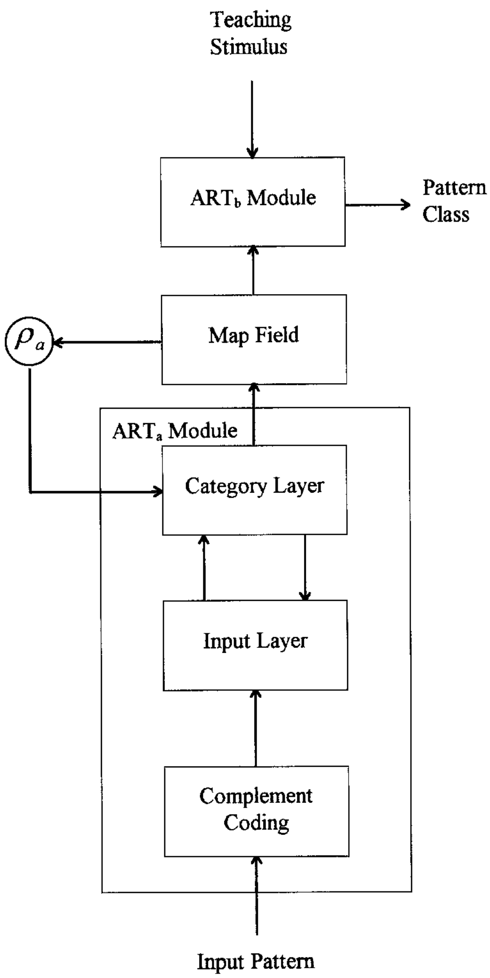

Fig. 3. Simplified fuzzy ARTMAP.

equal to zero and the learning rate, b, is set to its maximum value of one-so-called fast learning.) The only remaining user-changeable parameter is the baseline vigilance for the ART$_a$ module, ρ_a. This determines how close a match is required between an ART$_a$ input pattern and a category cluster prototype before accepting the input as a member of the cluster. This parameter (indirectly) controls the size of the category clusters that will form, since the higher it is set, the closer acceptable matches must be, and the smaller the coverage of the state space each cluster will have. Generally, higher vigilance provides better classification performance, although this must be balanced against the potential proliferation of category clusters, providing poor data compression and leading the network to become little more than a 'look-up table' [35]. Additionally, with small training sets and/or high-dimensional input vectors with many features, high vigilance can lead to incomplete coverage of the feature space by the network.

As well as its capabilities for continuous learning and symbolic rule extraction, SFAM has a number of other useful properties for medical pattern classification tasks: first, as noted earlier, the model has only one user-changeable parameter, the baseline vigilance of the ART$_a$ module. SFAM can thus be easily tuned to a particular task. Second, successful learning can occur with only one pass through the data set (termed single-epoch training). This is demonstrated within this paper, since all the results we describe were achieved by means of single-epoch training. Third, the model does not perform optimisation of an objective function and is not therefore prone to the problem of local minima as occurs with feedforward networks using backpropagation. Also, the problem of selecting the appropriate number of hidden units does not occur. This is because, as described previously, SFAM self-organises its own structuring of the data, automatically creating new category clusters for itself as and when they become needed. Fourth, the model is able to discriminate rare events from a 'sea' of similar cases with different outcomes owing to the feedback mechanism based on top–down matching of learned categories to input patterns. This is again in contrast to feedforward networks using backpropagation where weights are refined by a process which effectively averages together similar cases and thus fails to acknowledge rare events. SFAM is therefore suitable for domains where the distribution of data items is skewed between different categories. This is demonstrated within the present application domain (see Section 4.1).

Additionally, there are a number of supplementary features of our approach which require some explication.

3.4. Voting Strategy

The formation of category clusters in ARTMAP models is affected by the order of presentation of input data items [33]. Thus, the same data presented in a different order to separate SFAM networks can lead to the formation of quite different clusters within the two networks. This subsequently leads to different categorisations of test data, and thus different performance scores. This effect is particularly marked with small training sets and/or high-dimensional input vectors, where the input items may not be fully representative of the domain, and with single-epoch training. Sensitivity to order of presentation is not a problem of adaptive resonance theory alone. Indeed, in humans, the well known psychological effects of primacy and recency reflect precisely this ordering dependence. Rather the effect results from causal (on-line) learning where data enters the problem as it appears in nature, and its order is outwith the control of the experimenter. It is thus important in on-line learning to consider the ensemble of all possible learning experiments to understand the underlying behaviour of the system. Such effects are hidden when developing feedforward networks off-line because an artificial ensemble is created by multiple randomised data presentations during training. On-line training of an MLP may well lead to a different solution from one trained on a different ordering of the same data.

Order dependence can be compensated for by the use of a voting strategy [33]. This works as follows: a number of SFAM networks are trained on different orderings of the training data. During testing, each individual network makes its prediction for a test item in the normal way. The number of predictions made for each category is then totalled and the one with the highest score (or the most 'votes') is the final predicted category outcome. The voting strategy can provide improved SFAM performance in comparison with the individual networks. Furthermore, other voting methods can be used (see Section 3.4).

3.5. Symbolic Rule Extraction

Most neural networks suffer from the opaqueness of their learned associations [36]. In medical domains, this 'black box' nature may make clinicians reluctant to utilise a neural network application, no matter how great the claims made for its performance. Thus, there is a need to supplement neural networks with symbolic rule extraction capabilities to provide explanatory facilities for the network's 'reasoning'. The ARTMAP models have been endowed with such capabilities [37]. The act of rule extraction is a straightforward procedure compared with that required for feedforward networks since there are no hidden units with implicit meaning. In essence, each category cluster in ART_a represents a symbolic rule whose antecedent is the category prototype weights and whose consequent is the associated ART_b category (denoted via the map field).

3.6. Category Pruning

An SFAM network often becomes 'over-specified' on the training set, generating many low-utility ART_a category clusters which may represent noise or rare but unimportant cases, and subsequently provide poor-quality rules. The problem is particularly acute when a high ART_a baseline vigilance level is used during training. To overcome this difficulty, rule extraction involves a 'pre-processing' stage of category pruning. This involves the deletion of these low utility nodes.

Pruning is guided by the calculation of a Confidence Factor (CF) between nought and one for each category cluster, based upon a node's usage and accuracy. The usage score for an ART_a node is simply the number of training set exemplars it encodes, normalised through division by the maximum number of exemplars encoded by any node with the same category outcome. (Hence, there will be at least one node for each different category class which has a maximal usage score of one.) The accuracy score for a node is calculated as the proportion of predictions that are correct which the node mades on a prediction data set independent of the training data. This score is then normalised, similarly to the usage calculation, through division by the maximum proportion of correct predictions made by any node with the same outcome. (Thus, there will be at least one node for every category class which has a maximal accuracy score of one). The confidence factor for a node is then calculated as the mean of its usage and accuracy scores. All nodes with a confidence factor below a user-set threshold will be pruned. Full details of the process are given in Carpenter and Tan [37].

The pruning process can provide significant reductions in the size of a network. In addition, it also has the very useful side-effect that a pruned network's performance is usually superior to the original, unpruned net on both the prediction set and on entirely novel test data. This is because the removal of the low-utility nodes caused by over-specification on the training set improves the general performance of the network.

In the original formulation of the pruning process, a uniform CF threshold is used to select nodes for deletion, irrespective of their category class. In this application, we have generalised the pruning process to allow separate CF thresholds for nodes belonging to different category classes. This allows us to vary the proportion of the state-space covered by different categories. For example, by increasing the CF threshold for nodes with positive outcomes the relative proportion of such nodes is decreased and thus the sensitivity of the network is reduced. (The same effect can also be achieved of course by decreasing the CF threshold for nodes with negative outcomes.) This is useful for medical domains since it allows an ARTMAP network to be pruned so as to trade sensitivity for specificity and vice versa.

3.7. *SFAM* Cascade

The generalisation of the category pruning process described above allows a novel 'cascaded' configuration of SFAM networks to be employed as shown in Fig. 4. This consists of three layers, a network pruned so as to maximise sensitivity, another pruned so as to maximise specificity, and a third layer consisting of a set of uniformly pruned networks (which maximise accuracy) operating by the voting strategy. (The network cascade described here is a slightly simplified version of that first presented in Downs et al. [38] which was used there for the early diagnosis of heart attacks.)

The first two layers are intended to identify those cases which have a very high certainty of being classified correctly, with the sensitive network being used to 'trap' the negative cases and the specific network capturing the positive cases. The intuition behind this is that a network which displays very high sensitivity will rarely make false negative predictions and so any negative predictions made by that network are very likely to be correct. Conversely, a highly specific network will make very few false positive predictions, and so its positive predictions have a high certainty of being correct.

The cascade, therefore, operates as follows: An input data item is first presented to the sensitive network. If this yields a benign (negative) verdict, this is taken as the final category prediction. If not, the data item is next presented to the specific network. If this yields a malignant (positive) verdict, this is taken as the ultimate category prediction. Otherwise the final prediction of the category class of the input is obtained by majority verdict from the uniformly pruned nets, with a lower certainty of the prediction being correct than with the previous two layers. The cascade mimics the behaviour of a human pathologist in this domain in that some diagnoses are reported as 'certainly' correct, while others are labelled as 'suspicious' (and will thus require a second opinion and/or further test samples).

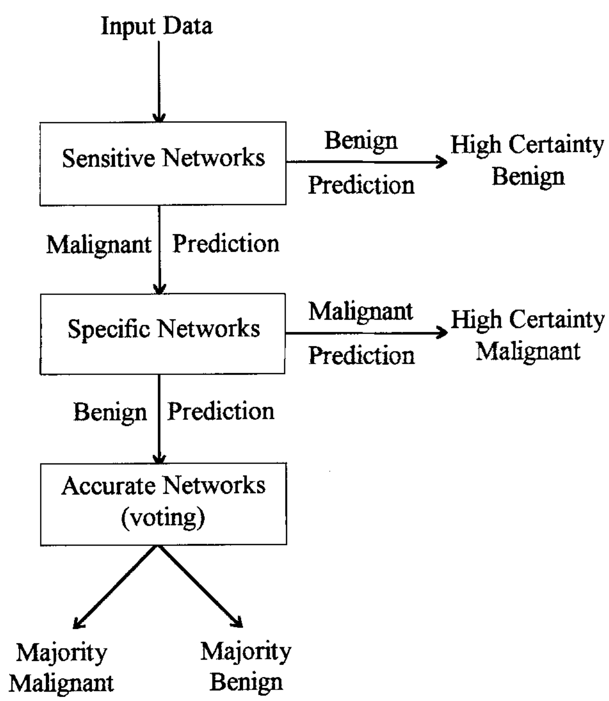

Fig. 4. Cascaded SFAM configuration for breast cancer diagnosis.

4. Basic Findings

4.1. Data

The total data set for the application comprised 600 patient records each consisting of 10 binary-valued features (see appendix). Each record was derived from microscopic observation of FNAB specimens by an expert pathologist (of Consultant status with 10 years experience in the field). The samples were taken from patients referred to the Royal Hallamshire Hospital, Sheffield, UK with symptomatic and screening-detected breast lesions in 1993. The distribution of categories within the data was slightly skewed but represents the prior probabilities of adequate specimens received in the laboratory – 215 cases were malignant, the remaining 385 benign (i.e. 35.8% malignant, 64.2% benign). This data differs therefore from that used in our previous work [27, 28], in that the previous data set was artificially biased to represent approximately equal numbers of each outcome, rather than the actual distribution across outcomes within the domain. (Additionally, the extra data in the current work allows performance to be measured on a test set that is independent of the prediction data set used to guide category pruning.)

In many medical situations, the final diagnosis or outcome is difficult to confirm without unnecessary invasive procedures or a long period of time has to elapse to allow further manifestations of the disease to appear. In FNAB, the final outcome is relatively easy to confirm within a few months of the initial procedure. In this study, the majority of cases where a malignant diagnosis was made on FNAB had further excision of tissue (e.g. lumpectomy or mastectomy) and the final diagnosis was made by histological examination which has a specificity and a sensitivity very close to 100%. In a few elderly patients, the tumour was not removed and the patient was treated by radiotherapy and chemotherapy, in these cases the clinical diagnosis of malignancy was secure since all were large tumours with clinical features of malignancy such as invasion through the skin. A benign outcome was confirmed by benign clinical features recorded at the time of aspiration, mammographic examination and absence of subsequent malignant specimens from the same area of the breast.

It should also be noted that, as with almost all information gathered from a real-world medical domain, the data set possesses a degree of 'noise'. Specifically, some combinations of features do not always have the same outcome in every case (notably owing to flawed aspirations of malignant lesions). Analysis of the data set revealed the existence of 16 such states, which collectively account for 365 cases. Assuming that the most frequent outcome should always be chosen when an ambiguous feature-state occurs will result in 22 of these cases being misclassified. This represents 3.7% of the total data-set, and, thus, the maximum possible diagnostic accuracy with this data is 96.3%.

4.2. Method

The total data were partitioned into three subsets; 125 randomly selected items formed a prediction set of 44 malignant and 81 benign cases (i.e. 35.2% malignant, 64.8% benign), a further 100 randomly selected items formed a test set of 35 malignant and 65 benign cases, the remaining 375 items formed the training set of 136 malignant and 239 benign cases (i.e. 36.3% malignant and 63.7% benign).

Ten SFAM networks were then trained on different random orderings of the teaching data. Vigilance was set very high (0.9) during training in order to maximise classification performance. Vigilance was relaxed to 0.6 for all predictions to ensure that all cases were matched to an existing category cluster node (i.e. forced choice prediction). Performance on the prediction set was recorded for each network in order to calculate the accuracy ratings for each category node as a prerequisite to category pruning.

The ten trained networks were then pruned in three separate ways. First, the 'standard' form of category pruning [37] was performed on the original networks, such that all nodes with a CF below 0.5 were deleted from the networks to improve predictive accuracy. The original networks were then pruned using different CF thresholds for the malignant and benign nodes to produce pruned networks which maximised sensitivity. CF thresholds of 0.05 for malignant nodes and 0.95 for benign nodes were employed, the criterion for setting the CF thresholds being a performance on the prediction set such that no network made more than one false negative prediction. A similar procedure was then conducted to produce 10 networks which maximised specificity. CF thresholds of 0.65 for malignant and 0.5 for benign nodes were sufficient to yield no more than one false positive for all networks on the prediction set.

The cascade configuration shown in Fig. 4 was then derived from the pruned networks. Selection criteria were as follows: From the ten networks pruned for sensitivity, the one which had the highest specificity while maintaining 100% sensitivity on the prediction set was selected to form the first layer of the cascade. Similarly, from the 10 networks pruned for specificity, the one which had the highest sensitivity while maintaining 100% specificity on the prediction set was chosen to form the second layer of the cascade. Finally, from the uniformly pruned networks, the five with the highest individual accuracies on the prediction set were selected to form the third (voting strategy) layer of the cascade. Performance of the cascade was then measured on the test set.

4.3. Results

The performance of the network cascade on the test set is shown in Table 1, alongside the performance of the human consultant on the same set of cases. It can be seen that the overall performance of the cascade is very similar to that of the human pathologist – in terms of numbers of cases, the cascade made two false positive predictions avoided by the consultant, but also made only three false negative predictions, compared with six made by the consultant. Additionally, the two false positive decisions were confined to the 'suspicious' layer of the cascade. The overall accuracy of the cascade can also be seen to be very close to the maximum possible for the domain (see Section 4.1).

The top two layers of the cascade (representing 'certain' decisions) showed almost perfect performance, and covered a large proportion of the total test data (71%). Perfect performance was marred only by the occurrence of one false negative decision. Further examination of the details of this case revealed that the sample was very likely to have resulted from a flawed aspiration, which, as noted earlier, is an unavoidable feature of the domain.

Table 1. Performance of SFAM voting strategy cascade on 100 item test set.

	Accuracy (%)	Sensitivity (%)	Specificity (%)
'Certain' decisions	98.6	96.7	100.0
'Suspicious' decisions	86.2	60.0	91.6
Overall performance	95.0	91.4	96.9
Consultant's performance	94.0	82.9	100.0

It should also be noted that standard category pruning produced a mean reduction in network size of 70.3%, from a mean of 91 category nodes in the unpruned networks to an average of 27 nodes in the pruned networks.

5. Symbolic Rules

5.1. General Discussion

Rule extraction facilities provide two advantages which, taken collectively, should help to overcome reluctance to utilise a neural network decision-support tool. First, a domain expert can examine the complete rule set in order to validate that the network has acquired an appropriate mapping of input features to category classes. Second, the symbolic rules provide explanatory facilities for the network's predictions during on-line operation. In the case of SFAM this corresponds to displaying the equivalent rule for the ART_a cluster node that was activated to provide a category decision. (In the case of the voting strategy, a number of such rules, one per voting network, would be displayed.) The diagnosing clinician is then able to decide whether or not to concur with the network's prediction, based upon how valid he or she believes the rule(s) to be.

The specific rules discovered for this domain will be presented in Section 5.2. However, some discussion of their general nature is needed here since they differ somewhat from the production rules used in conventional decision support aids such as expert systems. Expert system rules are 'hard' – an input must match to each and every feature in a rule's antecedent before the consequent will be asserted. In ARTMAP models, the rules are 'soft' – recall that they are derived from prototypical category clusters which are in competition with each other to match to the input data. Exact matching between inputs and categories is not necessary; a reasonably close fit suffices. (The degree of inexactitude that is tolerated being determined by the value of the ART_a vigilance parameter.) This provides greater coverage of the state space for the domain, using fewer rules.

Additionally, the rules are self-discovered through exposure to domain exemplars, rather than having been externally provided by a human expert. ARTMAP models are thus able to bypass the difficult and time-consuming knowledge-acquisition process found with rule-based expert systems [1]. However, collection of the data may itself be a non-trivial task in many medical domains.

A drawback of this approach is that the rules are 'correlational' rather than causal, since SFAM possesses no underlying theory of the domain but simply associates conjunctions of input features with category classes. (Of course, this problem is not specific to the ARTMAP models but occurs with neural networks generally.) However, this difficulty is probably not of great importance from an applications viewpoint since useful diagnostic performance can often be achieved from correlational features without recourse to any 'deep' knowledge of the domain.

A final general point concerns the learning rule in SFAM which governs the formation of category clusters, and hence the rules that will be derived from these clusters. Under the 'fast-learning' conditions with binary data used in this application, whenever an input is successfully matched to an existing category cluster node the new weights for that node are formed by taking the logical AND of the input pattern and the existing weights for that cluster [33]. This has the effect of deleting all features from the category cluster weights that are not also present in the input pattern. Hence, the weights tend to denote progressively more general clusters as they encode more input patterns and more features are deleted. Additionally, all features that are still present in the weights for a cluster once training ceases are known to have been present in all input vectors encoded by that cluster.

5.2. *SFAM* Rules for the Domain

Rule extraction from the uniformly pruned networks yielded 44 distinct rules, 12 with benign outcomes

Table 2. Benign rules extracted from uniformly pruned SFAM networks.

Rule 1 5 Occurrences Mean CF = 1.00 IF NOSYMPTOMS THEN BENIGN	Rule 2 5 Occurrences Mean CF = 0.57 IF NAKED THEN BENIGN	**Rule 3** **5 Occurrences** **Mean CF = 0.52** **IF** **FOAMY** **NUCLEOLI** **THEN** **BENIGN**	Rule 4 4 Occurrences Mean CF = 0.53 IF APOCRINE THEN BENIGN
Rule 5 **4 Occurrences** **Mean CF = 0.50** **IF** **PLEOMORPH** **THEN** **BENIGN**	**Rule 6** **4 Occurrences** **Mean CF = 0.50** **IF** **3D** **FOAMY** **THEN** **BENIGN**	Rule 7 3 Occurrences Mean CF = 0.52 IF FOAMY APOCRINE THEN BENIGN	Rule 8 3 Occurrences Mean CF = 0.52 IF PLEOMORPH SIZE APOCRINE THEN BENIGN
Rule 9 2 Occurrences Mean CF = 0.77 IF FOAMY THEN BENIGN	Rule 10 2 Occurrences Mean CF = 0.53 IF NAKED FOAMY THEN BENIGN	Rule 11 2 Occurrences Mean CF = 0.52 IF FOAMY PLEOMORPH SIZE APOCRINE THEN BENIGN	**Rule 12** **2 Occurrences** **Mean CF = 0.50** **IF** **3D** **NAKED** **NECROTIC** **THEN** **BENIGN**

(shown in Table 2) and 32 with malignant outcomes (shown in Table 3). Rules are ranked within each table by number of occurrences within the five networks and by the mean value of their certainty factor. The rules which the Consultant pathologist involved in this study did not consider to be in accord with standard diagnostic criteria are shown in bold within the tables.

Of the 44 rules extracted from the trained SFAM networks, 39 were in complete agreement with published canonical lists [24]. The majority of the rules for a malignant diagnosis consisted of combinations of the features ICL, 3D, NUCLEOLI, PLEOMORPH and SIZE with other features appearing less frequently. Most cytopathologists agree that the nuclear features of increased nuclear size (SIZE), multiple prominent nucleoli (NUCLEOLI) and variation in nuclear size (PLEOMORPH) are the most important diagnostic features of malignancy in FNAB [24,39,40]. Rule 17 for malignant diagnoses is interesting in that it only contains one feature, the presence of intracytoplasmic lumina; one specific type of breast carcinoma, lobular carcinoma, often produces cells which do not have prominent nuclear abnormalities but which do have these abnormal cytoplasmic inclusions [41]. Rule 29 for malignant diagnoses gives only the feature of increased nuclear size but most cytopathologists require other nuclear abnormalities to be present before making an unequivocally malignant diagnosis, with increased nuclear size alone a cytopathologist would probably assign the case to a suspicious category. In the benign rules, the features PLEOMORPH and SIZE appeared together twice which taken alone would be suggestive of malignancy but in these rules they were in combination with the feature APOCRINE. Apocrine change in breast epithelial cells produces large cells with abundant cytoplasm but the nucleus is also enlarged which would produce positive observations of the features PLEOMORPH and SIZE without prominent multiple nucleoli (NUCLEOLI). Apocrine change could also have led to the production of benign rule 5 where PLEOMORPH is the only feature since a few cells with apocrine change in the specimen would trigger this positive observation without also triggering APOCRINE because the majority of epithelial cells did not show apocrine change. This circumstance arises because of the binary nature of the definitions used in this study, a coding scheme with more gradation for

Table 3. Malignant rules extracted from uniformly pruned SFAM networks.

Rule 1 5 Occurrences Mean CF = 0.94 IF 3D NUCLEOLI PLEOMORPH SIZE THEN MALIGNANT	Rule 2 5 Occurrences Mean CF = 0.60 IF DYS ICL NUCLEOLI PLEOMORPH SIZE NECROTIC THEN MALIGNANT	Rule 3 5 Occurrences Mean CF = 0.59 IF 3D NUCLEOLI PLEOMORPH SIZE NECROTIC THEN MALIGNANT	Rule 4 4 Occurrences Mean CF = 0.99 IF ICL 3D NUCLEOLI PLEOMORPH SIZE THEN MALIGNANT
Rule 5 4 Occurrences Mean CF = 0.75 IF DYS NUCLEOLI PLEOMORPH SIZE THEN MALIGNANT	Rule 6 4 Occurrences Mean CF = 0.73 IF NUCLEOLI PLEOMORPH SIZE THEN MALIGNANT	Rule 7 4 Occurrences Mean CF = 0.64 IF PLEOMORPH SIZE = TRUE THEN MALIGNANT	Rule 8 4 Occurrences Mean CF = 0.59 IF ICL 3D FOAMY NUCLEOLI PLEOMORPH SIZE NECROTIC THEN MALIGNANT
Rule 9 4 Occurrences Mean CF = 0.58 IF ICL FOAMY PLEOMORPH SIZE THEN MALIGNANT	Rule 10 4 Occurrences Mean CF = 0.58 IF ICL NUCLEOLI PLEOMORPH SIZE THEN MALIGNANT	Rule 11 4 Occurrences Mean CF = 0.56 IF DYS 3D FOAMY NUCLEOLI PLEOMORPH SIZE NECROTIC THEN MALIGNANT	Rule 12 4 Occurrences Mean CF = 0.55 IF 3D NAKED NUCLEOLI PLEOMORPH SIZE THEN MALIGNANT
Rule 13 3 Occurrences Mean CF = 0.79 IF 3D FOAMY NUCLEOLI PLEOMORPH SIZE THEN MALIGNANT	Rule 14 3 Occurrences Mean CF = 0.69 IF FOAMY NUCLEOLI PLEOMORPH SIZE THEN MALIGNANT	Rule 15 3 Occurrences Mean CF = 0.68 IF FOAMY NUCLEOLI PLEOMORPH SIZE NECROTIC THEN MALIGNANT	Rule 16 3 Occurrences Mean CF = 0.62 IF 3D FOAMY PLEOMORPH SIZE NECROTIC THEN MALIGNANT
Rule 17 3 Occurrences Mean CF = 0.60 IF ICL THEN MALIGNANT	Rule 18 3 Occurrences Mean CF = 0.57 IF ICL 3D NUCLEOLI PLEOMORPH SIZE NECROTIC THEN MALIGNANT	Rule 19 3 Occurrences Mean CF = 0.53 IF ICL NAKED FOAMY NUCLEOLI PLEOMORPH SIZE THEN MALIGNANT	Rule 20 2 Occurrences Mean CF = 0.67 IF ICL 3D NUCLEOLI THEN MALIGNANT

Table 3. Continued.

Rule 21 2 Occurrences Mean CF = 0.62 IF DYS ICL NUCLEOLI PLEOMORPH SIZE = TRUE THEN MALIGNANT	Rule 22 2 Occurrences Mean CF = 0.62 IF NUCLEOLI PLEOMORPH SIZE NECROTIC THEN MALIGNANT	Rule 23 2 Occurrences Mean CF = 0.62 IF ICL FOAMY NUCLEOLI PLEOMORPH SIZE THEN MALIGNANT	Rule 24 2 Occurrences Mean CF = 0.60 IF DYS NUCLEOLI PLEOMORPH SIZE NECROTIC THEN MALIGNANT
Rule 25 2 Occurrences Mean CF = 0.57 IF ICL PLEOMORPH SIZE THEN MALIGNANT	Rule 26 2 Occurrences Mean CF = 0.59 IF ICL FOAMY THEN MALIGNANT	Rule 27 2 Occurrences Mean CF = 0.55 IF DYS FOAMY NUCLEOLI PLEOMORPH SIZE NECROTIC THEN MALIGNANT	Rule 28 2 Occurrences Mean CF = 0.53 IF DYS ICL 3D NUCLEOLI PLEOMORPH SIZE THEN MALIGNANT
Rule 29 **1 Occurrence** **Mean CF = 0.63** **IF** **SIZE** **THEN** **MALIGNANT**	Rule 30 1 Occurrence Mean CF = 0.59 IF DYS 3D NUCLEOLI PLEOMORPH SIZE THEN MALIGNANT	Rule 31 1 Occurrence Mean CF = 0.56 IF ICL 3D NAKED FOAMY NUCLEOLI SIZE THEN MALIGNANT	Rule 32 1 Occurrence Mean CF = 0.53 IF DYS ICL 3D NUCLEOLI PLEOMORPH SIZE NECROTIC THEN MALIGNANT

each feature could circumvent this problem. Benign rules 3, 6 and 12 are not in complete accord with published cytopathological knowledge and these cases would probably have been assigned to a suspicious category by a cytopathologist.

6. Further Results

This section presents results concerning investigations into the effects upon SFAM performance of employing both different input features and different observers to derive the features. It follows up issues raised in Downs et al. [28].

6.1. Effect of Different Input Features

In this section, we consider three variations of the 'standard' set of input features used to obtain the results described in Section 4:

1. The removal of the 'FOAMY' feature (representing the presence of foamy macrophages in the background of the slide). In previous work [28], it was observed that there was disagreement as to the utility of this feature in making a diagnosis. This variation therefore attempts to resolve this issue by determining if removal of the 'FOAMY' feature has an adverse effect upon network performance.
2. The addition of an 'AGE' feature, representing the patient's age. (The age feature was encoded for input to SFAM by dividing the patient age in years by 100 in order to yield a value in the range 0–1.) The motive for including this feature is that after menopause the breast atrophies, and thus the proportion of epithelial cells drops. Aspirates are expected therefore to show fewer and smaller epithelial cells. Hence, patient age is a potentially useful diagnostic indicator.
3. The addition of a binary 'NOISE' feature, derived

from whether or not the laboratory number of the sample was odd or even. This variation is intended to investigate the robustness of SFAM in the face of irrelevant data.

To compare the effect of different input features, the SFAM network configuration and experimental method were changed from the procedure described in Section 4.2. This earlier method was intended to optimise SFAM performance for use of the model as a decision-support tool. This goal was achieved, with a diagnostic accuracy very close to the maximum possible for the domain. However, comparison of the effects of different input features requires a non-optimal baseline performance in order to detect any improvements caused by input feature variations. The following changes in the experimental method were therefore adopted, all for the purpose of lowering the baseline SFAM performance with the standard set of input features:

1. Category pruning was omitted. This also allowed an expanded test set to be created, by merging the prediction and test sets described in Section 4.2 into a single 225 item test set.
2. Vigilance was set at zero during training of the networks. This change was further necessitated by the fact that a uniform non-zero vigilance level can have a slightly different effect upon two SFAM networks which have different sized input vectors [42].
3. The full SFAM cascade was not employed in measuring performance. Instead, a simpler three network voting strategy was used.

The full description of the experimental method is as follows. Ten SFAM networks were trained on different random orderings of the 375 item teaching data set described in Section 4.2, using single-epoch training with a vigilance level of zero. These networks used the 10 'baseline' standard input features described previously in the appendix. Performance of each individual network was then recorded on the new 225 item test set. The three networks with the highest individual accuracy were then selected to form a three network voting strategy. The voting strategy performance of these networks was then recorded on the test set.

The procedure described in the above paragraph was then repeated using the three variations in the network input features outlined previously. SFAM results with the three input feature variations were then compared with the baseline results across all performance metrics (i.e. accuracy, sensitivity and specificity) using McNemar's test [43].

The performance with the three input feature variations in comparison to the standard input feature set is shown in Table 4. It can be seen that the deletion of the 'FOAMY' feature led to a small drop in overall accuracy because of reduced specificity. However, none of the differences across the three metrics was statistically significant (McNemar's test, $p > 0.1$). The addition of the 'AGE' feature led to an increase across all performance metrics in comparison to the baseline. These increases were all statistically significant (McNemar's test; accuracy $p < 0.001$, sensitivity $p < 0.05$, specificity $p < 0.01$). The addition of the 'NOISE' feature did not cause a deterioration in performance in comparison to the baseline. Indeed, it led to a slight increase in accuracy owing to higher specificity. However, neither difference was statistically significant (McNemar's test, $p > 0.1$).

6.2. Effect of Different Observers

In previous work [28], performance of the SFAM network appeared to degrade badly in the face of 'noisy' input data provided by an inexperienced junior pathologist. It is thus possible that SFAM's utility as a decision-support tool is limited by the quality of the input data, being suitable for use by senior pathologists but vulnerable in the face of incorrect feature assignments made by junior pathologists. In this subsection therefore, this issue is investigated further by studying the performance of SFAM using input feature observations provided by several different pathologists.

A test set of 50 cases (25 malignant and 25 benign) was employed which was entirely inde-

Table 4. Effect of different input features on SFAM performance.

Input features	Accuracy (%)	Sensitivity (%)	Specificity (%)
BASELINE 10	85.3	84.8	85.6
–FOAMY	82.2	87.3	79.5
+AGE	94.2	94.9	93.8
+NOISE	87.6	84.8	89.0

pendent of the 600 item data set described previously in Section 4.1. The presence or absence of the ten standard input features (see the Appendix) for each case was then rated by six different pathologists, all working separately and with no knowledge of clinical data apart from patient age. One of these pathologists had provided the original 600 item data set. Of the six pathologists, two were of consultant status (each with over 10 years experience of interpreting FNAB), three were senior registrars (each with six years experience) and one a senior house officer (two years experience). All pathologists reported FNAB in their daily work but those below consultant status would have had their reports scrutinised by a consultant pathologist who might modify the report after examining the cytological preparations.

Voting strategy performance of the five uniformly pruned networks trained previously to form the third layer of the SFAM cascade (see Section 4) was recorded for each of the six pathologist's observations. The results with the observations from the pathologist who had provided the original training data served as the baseline for comparison with the other five pathologists' observations (observer 1 in Table 5). The results for all three performance metrics were compared using McNemar's test.

The performance of SFAM with the six sets of observations is shown in Table 5. It should be noted that the small size of the data set used here means that the differences between observers may appear somewhat exaggerated since, in percentage terms, a single false case alters specificity or sensitivity by 4% and accuracy by 2%.

General performance of the SFAM networks held up well across all observers, with no more than four misclassifications for all but one observer's data (seven misclassifications with observer 4). Somewhat surprisingly, the networks performed slightly better than the baseline with two observers (numbers 3 and 5). Furthermore, the observation set showing the worst performance had only four extra misclassifications in comparison to the baseline. There were no statistically significant differences between the baseline observation set and any of the other observation sets across any performance metric (McNemar's test, $p > 0.1$).

7. Discussion and Conclusions

The performance of the SFAM cascade described in Section 4.3 indicates that the architecture has useful potential value as a decision-support tool for this domain. Performance is very close to the domain optimum, despite only utilising single-epoch training on a skewed data set. Furthermore, the cascade served to isolate a large proportion of cases where the network's predictions were very likely to be correct, and, very importantly, avoided any false positives within this subset of the test data.

Additionally, the model was shown to have self-discovered a generally valid set of rules for the domain, with only five rules from a set of 44 considered not to be 'canonical'. It is interesting to note that all but one of these five rules had a mean CF factor very close to the threshold for pruning. We therefore believe it is possible that the networks were 'under-pruned' in this application and that a slightly higher CF threshold for pruning (e.g. 0.55) would have almost entirely eliminated all dubious rules, as well as providing a more compact overall rule set. However, balanced against this is the need to avoid 'over-pruning' where useful category nodes are deleted resulting in degraded system performance and incomplete coverage of the state space for the domain. This is therefore an area for future work. A further interesting related piece of future work might be to use expert knowledge as a 'post-processor' for pruning, by deleting only those rules above the standard pruning threshold which are considered to be invalid by the domain expert, and recording any subsequent effects upon system performance.

The rule discovery aspect of SFAM can be seen as a knowledge engineering facet of the model – confirming that a network behaves in a way that is acceptable to a domain expert in order to enhance

Table 5. Effect of different observers on uniformly pruned SFAM performance.

Observer	Accuracy (%)	Sensitivity (%)	Specificity (%)
1	94	100	88
2	92	92	92
3	96	100	92
4	86	84	88
5	96	100	92
6	94	92	96

confidence in the use of the model as a decision-support tool. Conversely, our experiments with variations in the input features illustrate a machine learning facet-establishing the utility of different diagnostic features which may be at odds with the 'received wisdom' of domain experts. Specifically, our results indicate that patient age should be regarded as a useful diagnostic feature, whereas the presence of foamy macrophages is not perhaps as useful as is conventionally claimed. Furthermore, the SFAM was shown to be robust in the face of an irrelevant input feature which did not degrade system performance (see also Goodman et al. [44] for another experiment concerning fuzzy ART-MAP's noise resistance in a different medical domain – the prediction of the length of hospital stay of pneumonia patients).

Our findings with different observers providing the network inputs also showed the SFAM network to be quite robust across potential variations caused by subjectivity in feature assignments made by different pathologists. This is somewhat at odds with our previous findings [28]. The previous study, which showed a marked decrease in performance with an inexperienced observer, only used two observers and the inexperienced observer had 18 months experience compared with a minimum of two years in this study. This is therefore an area where more research is needed in order to establish the limits of and reasons for SFAM's apparently variable robustness, probably requiring a larger volume of data providing feature ratings from a variety of pathologists of different levels of experience. The aptitude of individuals for pattern recognition and visual discrimination tasks varies and larger numbers of inexperienced observers are required before any firm conclusions about the system's performance with very junior pathologists can be made. Unfortunately, the collection of data for such a study is a non-trivial task. Evidently, an objective method for feature extraction would enhance the utility of the current system and this remains an area for future work.

A problem with studies where human observers are used to make observations which are then used in a decision-support system is the separation of observational and interpretative processes. In the cytodiagnosis of FNAB in a routine (non-teaching) setting, pathologists rarely express their decisions in component observations and the diagnosis is made by a process which is largely subconscious and is the result of training the natural neural networks of the human brain by observation of past specimens and knowledge of their diagnosis/outcome. Since the pathologists in this study knew the 'canonical' lists of features for the diagnosis of FNAB it is possible that their observations may have been biased if they made a diagnosis of the specimen whilst recording the observed features. The presence of five extracted rules which do not correspond with the 'canonical' lists is evidence against this as is the evidence that the SFAM voting strategy cascade produced a better overall performance than the human who made the observations which the SFAMs used. The separation between observations and diagnosis could be improved by selecting microscopic fields of view in specimens, digitising them and then displaying them to observers in random order so that fields from the same specimen were not in an unbroken sequence, but this would require numerous images for each specimen (since some of the definitions are based on the presence or absence of a feature in all cells of the specimen) and the resolution of the best digitised images is still not as great as those viewed directly through the microscope.

In this study, verbal definitions of observed features were used and the observers were just given these with no additional training and no sample images. The reliability of observations, especially among trainee pathologists, could be improved by a more visually-based input system such as a range of sample images for each feature displayed on a computer screen with a cursor controlled input scale [45].

We are also considering the development of a modified version of SFAM with a more sophisticated matching technique between input cases and category clusters. In SFAM, each true input feature contributes equally to the match with a category prototype. We envisage introducing a variable weighting for features, which attaches more importance to individual features that are considered to be (a) very strongly predictive for the domain and (b) most easily identified by an inexperienced pathologist. This should increase the robustness of an SFAM-based decision-support tool when used by an inexperienced pathologist, by effectively building *a priori* expert knowledge into SFAM.

Another area for future work is to automate the CF threshold selection process for the differential category pruning described in Section 3.6. In the present implementation, the CF thresholds were 'hand-set' by the system's designer to achieve the desired changes in network performance. However, this is a laborious trial-and-error process which contrasts poorly with the general ease of tuning of the basic SFAM model.

Finally, an important area for future work is to investigate the claimed potential for incremental learning with the SFAM model. The relatively small size of the data set did not allow such a study to be made in this application. However, were further data for the domain to become available, the existing networks could be trained on this new data without the necessity also to retrain with the original data. It should be noted though that there are a few caveats to this statement. Specifically, the voting strategy and category pruning are essentially 'off-line' learning processes which cannot immediately be employed in conjunction with continuous (case-by-case) learning. The voting strategy requires randomisation of the ordering of input data which obviously disrupts its original temporal order. Similarly, the category pruning process requires a 'batch' of input data to form the prediction set. However, these features can still be feasibly employed in the less stringent circumstance of incremental learning on a batch of new data, rather than on-line learning on a case-by-case basis. For example, suppose a further 125 of the most recent hospital cases were to become available for the domain. This could form a new prediction set and the existing 125 item prediction set would be freed to serve as further training data. This new training data could then be randomised and applied to the ten existing unpruned networks. Category pruning could then occur on the basis of performance on the new prediction set. This would yield new pruned networks to be employed in the SFAM cascade which should be adapted to the most recent data while still retaining useful older associations, without having to have undertaken extensive retraining from scratch.

Acknowledgements. This research was supported by the Engineering and Physical Sciences Research Council (EPSRC) of the UK, grant number GR/J/432333.

References

1. Hayes-Roth F, Waterman DA, Lenat DB. Building Expert Systems. Addison-Wesley, London, 1983
2. Bounds DG, Lloyd PJ, Mathew BG. A comparison of neural network and other pattern recognition approaches to the diagnosis of low back disorders. Neural Networks 1990; 3: 583–591
3. Egmont-Petersen M, Talmon JL, Brender J, McNair P. On the quality of neural network classifiers. Artif Intell in Medicine 1994; 6: 359–381
4. Harrison RF, Marshall SJ, Kennedy RL. A connectionist aid to the early diagnosis of myocardial infarction. Proc Third Euro Conf Artif Intell in Medicine, Maastricht, 1991: 119–128
5. Apolloni B, Avanzini G, Cesa-Bianci N, Ronchini G. Diagnosis of epilepsy via backpropagation. Proc Int Joint Conf Neural Networks 1990: 571–574
6. Pizzi N, Choo LP, Mansfield J, Halliday WC, Mantsch HH, Somorjal RL. Neural network classification of infrared spectra of control and Alzheimer's diseased tissue. Artif Intell in Medicine 1997; 7: 67–79
7. Cross SS, Harrison RF, Kennedy RL. Introduction to Neural networks. Lancet 1995; 346: 1075–1079
8. Baxt WG. Application of artificial neural networks to clinical medicine. Lancet 1995; 346: 1135–1138
9. Dybowski R, Gant V. Artificial neural networks in pathology and medical laboratories. Lancet 1995; 346: 1203–1207
10. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986; 323: 533–536
11. Moody J, Darken C. Fast learning in networks of locally-tuned processing units. Neural Computation 1989; 1: 281–294
12. Cybenko G. Approximations by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 1989; 2: 303–314
13. Park J, Sandberg I. Universal approximation using radial basis function networks. Neural Computation 1991; 3: 246–257
14. Richard M, Lippman R. Neural network classifiers estimate Bayesian a posteriori probabilities. Neural Computation 1991; 3: 461–483
15. Carpenter G, Grossberg S. The ART of adaptive pattern recognition by a self-organising neural network. Computer 1988; 21: 77–88
16. Sharkey NE, Sharkey AJC. An analysis of catastrophic interference. Connection Science 1995; 7: 301–329
17. Ma Z, Harrison RF. GR2: a hybrid knowledge-based system using general rules. Proc Int Joint Conf Artif Intell, Montreal, 1995: 488–493
18. Downs J, Harrison RF, Kennedy RL, Cross SS. Application of the fuzzy ARTMAP neural network model to medical pattern classification tasks. Artif Intell in Medicine 1996; 8: 403–428
19. Carpenter G, Grossberg S, Reynolds J. ARTMAP: Supervised real-time learning and classification of non-stationary data by a self-organizing neural network. Neural Networks 1991; 4: 565–588
20. Underwood JCE. Tumours: benign and malignant. In: Underwood JCE (ed), General and Systematic Pathology, Churchill Livingstone, Edinburgh, 1992, pp. 223–246
21. Elston CW, Ellis IO. Pathology and breast screening. Histopathology 1990; 16: 109–118
22. Wolberg WH, Mangasarian OL. Computer-designed expert systems for breast cytology diagnosis. Analytical and Quantitative Cytology and Histology 1993; 15: 67–74
23. Start RD, Silcocks PB, Cross SS, Smith JHF. Problems with audit of a new fine-needle aspiration service in a district general hospital. Pathology 1992; 167: 141A
24. Wells CA, Ellis IO, Zakhour HD, Wilson AR. Guidelines for cytology procedures and reporting on fine needle aspirates of the breast. Cytopathology 1994; 5: 316–334
25. Heathfield HA, Kirkham N, Ellis IO, Winstanley G. Computer assisted diagnosis of fine needle aspirates of the breast. J Clinical Pathology 1990; 43: 168–170
26. Hamilton PW, Anderson N, Bartels PH, Thompson D. Expert system support using Bayesian belief networks in the diagnosis of fine needle aspiration biopsy

specimens of the breast. J Clinical Pathology 1994; 47: 329–336
27. Downs J, Harrison RF, Cross SS. A neural network decision support tool for the diagnosis of breast cancer. In: Hallam J (ed), Hybrid Problems, Hybrid Solutions, IOS Press, Amsterdam, 1995, pp. 51–60
28. Downs J, Harrison RF, Cross SS. Evaluating a neural network decision-support tool for the diagnosis of breast cancer. Proc Fifth Euro Conf Artif Intell in Medicine (AIME-95), Pavia, Italy, 1995: 239–250
29. Carpenter G, Grossberg S. A massively parallel architecture for a self-organizing neural pattern recognition machine. Computer Vision, Graphics and Image Processing 1987; 37: 54–115
30. Grossberg S. Competitive learning: from interactive activation to adaptive resonance. Cognitive Sci 1987; 11: 23–63
31. Carpenter G. Distributed activation, search and learning by ART and ARTMAP neural networks. Research Report CAS/CNS-96-006, University of Boston, Boston, USA, 1996
32. Carpenter G, Grossberg S, Rosen D. Fuzzy ART: Fast, stable learning and categorisation of analogue patterns by an adaptive resonance system. Neural Networks 1991; 4: 759–771
33. Carpenter G, Grossberg S, Markuzon S, Reynolds J, Rosen D. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multi-dimensional maps. IEEE Trans Artif Neural Networks 1992; 3: 698–712
34. Kasuba T. Simplified fuzzy ARTMAP. AI Expert 1993; 8: 18–25
35. Marriott S, Harrison RF. A modified fuzzy ARTMAP architecture for the approximation of noisy mappings. Neural Networks 1995; 8: 619–642
36. Towell GG, Shavlik JW. Extracting refined rules from knowledge-based neural networks. Machine Learning 1993; 13: 71–101
37. Carpenter G, Tan A-H. Rule extraction, fuzzy ARTMAP and medical databases. Proc World Congress on Neural Networks 1993: 501–506
38. Downs J, Harrison RF, Kennedy RL. A prototype neural network decision-support tool for the diagnosis of acute myocardial infarction. Proc Fifth Euro Conf Artif Intell in Medicine (AIME-95), Pavia, Italy, 1995: 355–366
39. Trott PA. Aspiration cytodiagnosis of the breast. Diagnostic Oncology 1991; 1: 79–87
40. Bottles K, Chan JS, Holly EA, Chiu S, Miller TR. Cytologic criteria for fibroadenoma. Am J Clinical Pathology 1988; 89: 707–713
41. Quincey C, Raitt N, Bell J, Ellis IO. Intracytoplasmic lumina – a useful diagnostic feature of adenocarcinomas. Histopathology 1991; 19: 83–87
42. Stork DG. Self-organisation, pattern recognition and adaptive resonance networks. J Neural Network Comput 1989; 1: 26–42
43. Bland M. An Introduction to Medical Statistics, Oxford University Press, Oxford, 1987
44. Goodman PH, Kaburlasos VG, Egbert DD, Carpenter G, Grossberg S, Reynolds J, Rosen D et al. Fuzzy ARTMAP neural network compared to linear discriminant analysis prediction of the length of hospital stay in patients with pneumonia. In: Marks RJ (ed), Fuzzy Logic Technology and Applications. IEEE Technical Activities Board, Piscataway, 1994: 424–429
45. Hamilton PW, Bartels PH, Montironi R, Anderson N, Thompson D. Improved diagnostic decision-making in pathology: do inference networks hold the key? J Pathology 1995; 175: 1–6

Nomenclature

DYS:	True if majority of epithelial cells are dyhesive, false if majority of epithelial cells are in cohesive groups.
ICL:	True if intracytoplasmic lumina are present, false if absent.
3D:	True if some clusters of epithelial cells are not flat (more than two nuclei thick) and this is not due to artefactual folding, false if all clusters of epithelial cells are flat.
NAKED:	True if bipolar 'naked' nuclei in background, false if absent.
FOAMY:	True if 'foamy' macrophages present in background, false if absent.
NUCLEOLI:	True if more than three easily visible nucleoli in some epithelial cells, false if three or fewer easily visible nucleoli in epithelial cells.
PLEOMORPH:	True if some epithelial cell nuclei with diameters twice that of other epithelial cell nuclei, false if no epithelial cell nuclei twice the diameter of other epithelial cell nuclei.
SIZE:	True if some epithelial cells with nuclear diameters at least twice that of lymphocyte nuclei, false if all epithelial cell nuclei with nuclear diameters less than twice that of lymphocyte nuclei.
NECROTIC:	True if necrotic epithelial cells present, false if absent.
APOCRINE:	True if apocrine change present in all epithelial cells, false if not present in all epithelial cells.