ELSEVIER

Artificial Intelligence in Medicine 28 (2003) 165–189

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue

Matthias E. Futschik[a,*], Anthony Reeve[b], Nikola Kasabov[a]

[a]*Department of Information Science, University of Otago,*
P.O. Box 56, Dunedin, New Zealand
[b]*Department of Biochemistry, University of Otago, P.O. Box 56, Dunedin, New Zealand*

Received 26 October 2002; received in revised form 20 January 2003; accepted 14 April 2003

Abstract

Microarray techniques have made it possible to observe the expression of thousands of genes simultaneously. They have recently been applied to study gene expression patterns in tissue samples. This may lead to highly desirable improvements in the diagnosis and treatment of human diseases. Statistical and machine learning methods have recently been used to classify cancer tissue based on gene expression data. Although some of these methods have achieved a high degree of accuracy, they generally lack transparency in their classification process. This, however, is crucial for the application in the medical field. In order to overcome this obstacle, we used knowledge-based neurocomputing (KBN), since KBN seeks to gain knowledge that is comprehensible to humans. In particular, we applied evolving fuzzy neural networks (EFuNNs) to classify cancer tissue, which is illustrated on the case studies of leukaemia and colon cancer. EFuNNs belong to the evolving connectionist system paradigm (ECOS) that has been recently introduced. They are well suited for adaptive learning and knowledge discovery. Fuzzy logic rules can be extracted from the trained networks and offer knowledge about the classification process in an easily accessible form. These rules point to genes that are strongly associated with specific types of cancer and may be used for the development of new tests and treatment discoveries.

Keywords: Knowledge-based neural networks; Gene expression analysis; Evolving fuzzy neural networks; Rule extraction; Knowledge discovery

[*] Corresponding author. Tel.: +64-3-479-8316; fax: +64-3-479-8311.
E-mail addresses: mfutschik@infoscience.otago.ac.nz (M.E. Futschik), Anthony.Reeve@stonebow.otago.ac.nz (A. Reeve), nkasabov@otago.ac.nz (N. Kasabov).

0933-3657/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0933-3657(03)00063-0

1. Introduction

The recent advent of cDNA and oligonucleotide microarray technology means that it has now become possible to analyse thousands of genes in cell cultures and tissue samples simultaneously. The transcription activities of genes can be measured on a genome-wide scale by hybridisation of cDNA to oligonucleotides that have been printed on glass slides or attached to a glass surface by photolithography. The potential applications of these techniques are numerous. Regulatory networks can be discovered by analysis of time series experiments [20]. Fast functional assignment of novel genes can be achieved by clustering expression profiles because functionally related genes frequently show similar expression patterns. These gene expression clusters can be used to assign novel genes to well-known functions [7]. Besides the use of microarrays to explore gene regulation, the study of human diseases has become a major focus of recent research activities. Possible medical applications include the identification of markers for classification, diagnosis, disease outcome prediction, therapeutic responsiveness and target identification. Especially, cancer researchers hope to gain new insights into the development and characteristics of tumours [14]. Since the cause of cancer is generally linked to a complex interaction of several genes, only the simultaneous monitoring of many genes may reveal the underlying biological mechanisms.

Microarray studies of different kinds of cancer have shown that gene expression data can supplement previous methods of phenotyping of cancers which rely on a limited set of histological and pathological features. These traditional methods depend heavily on the experience of the physician to reach the correct decision. They are further limited since tumours with similar histopathological appearances can show very different responses to the same treatment. Any method that would facilitate the decision process for physicians is highly desirable. The microarray studies that have been published so far used different data analysis techniques, mainly for clustering and classification. The application of clustering methods demonstrated that it is possible to discover new subclasses of cancer [1]. The classification of cancer types was achieved by using statistical as well as machine learning techniques [8,13]. Altogether, microarray techniques have shown to be valuable and powerful candidates for enhancing decision making systems in the medical field.

A major problem, however, remains in the extraction of knowledge from the trained classification systems. Statistical methods are highly model dependent and restricted by prior assumptions about the data. Machine learning techniques, such as artificial neural networks (ANNs), present a more flexible 'model-free' approach for classification and frequently yield a good performance. As they are, however, 'black box' methods, they lack the power to 'explain' decision processes, which is necessary for any wide-spread application in the medical field. Crucial decisions about disease treatment demand that the classification process is transparent. Physicians need to understand how a classifier reaches its judgement, since the final responsibility for any decision in the course of a treatment remains with them. Furthermore, cancer researchers could profit from classifiers that can indicate in a comprehensive way which combinations of genes are indicative of certain cancer types. This may give new insights about the complex genetic interactions during the development of the disease.

These reasons motivated us to use knowledge-based neurocomputation (KBN) for microarray data analysis. KBN applies neural networks as a powerful computational model, while at the same time it seeks to overcome the drawbacks of the 'black box' approach [4]. KBN is an attractive paradigm for medical application, since it tries to construct classification systems which are able to represent the acquired knowledge in a way comprehensible to humans. KBN has been used for the analysis of DNA and RNA data [21], but it has not been employed for microarray data analysis so far.

In particular, we have applied evolving connectionist systems (ECOS) as a version of KBN [12]. ECOS are dynamically growing (and shrinking) neural networks for adaptive learning and rule elucidation from an input–output stream of data. For the analysis of gene expression data we used evolving fuzzy neural networks (EFuNNs) that are a specific implementation of ECOS. Based on fuzzy theory, EFuNNs allow for fuzzy rule extraction and adaptation [12]. Using fuzzy rules may be of advantage since microarray data contain a large noise component and crisp concepts are difficult to define at this stage of research.

While the usual focus of previous studies is set mainly on the performance of certain classifiers, we present in this paper a more holistic approach to tissue classification based on microarray data. We see the classification algorithm as an integral part of a larger information system for the analysis of microarray data. The main phases of information processing and problem solving in such systems are the following:

(1) Feature analysis and feature extraction—defining which features (e.g. genes) are more relevant and, therefore, should be used when creating a model for a particular problem (e.g. classification, prediction, decision making).
(2) Modeling the problem—consists of defining inputs, outputs and type of the model (e.g. probabilistic, rule-based, connectionist), training the model, etc.
(3) Knowledge discovery—new knowledge is gained through the analysis of the modeling results and the model itself.

To achieve a good performance, all these phases have to be integrated and optimised.

The structure of this paper is as follows: In the next section, we give a description of microarray experiments in general and the gene expression data sets that we used in our case studies. We discuss the major difficulties in the analysis of microarray data and develop a methodology for tissue classification based on gene expression data. This leads to the use of KBN, to which we give a brief introduction. We then describe the ECOS paradigm within the KBN framework and the EFuNNs as a specific version of ECOS. We outline the structure and the learning algorithm of EFuNNs and present an algorithm for rule extraction. This is followed by the application of the proposed methodology to the case study data sets of leukaemia and colon cancer. We conclude this study with a discussion on the performance analysis and describe the directions for our future research.

2. Microarray experiments and data flow

Microarrays have enabled researchers to monitor thousands of genes simultaneously. They consist of spatially ordered probes of cDNA or oligonucleotides on glass or nylon. These probes are selected to hybridise to complementary cDNA. The steps of a typical microarray experiment are shown in Fig. 1. The first step consists of RNA extraction from a tissue sample and eventual amplification. The RNA is reverse transcribed to cDNA that is

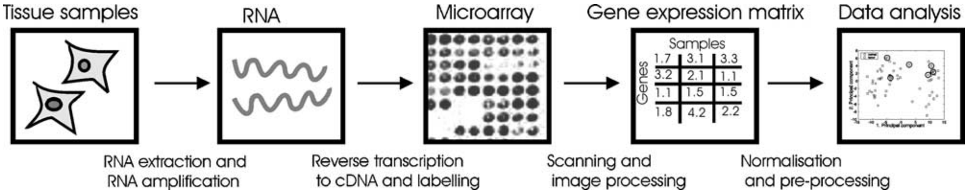
Tissue samples
RNA
Microarray
Gene expression matrix
Samples
Genes
1.7 3.1 3.3
3.2 2.1 1.1
1.1 1.5 1.5
1.8 4.2 2.2
Data analysis
RNA extraction and RNA amplification
Reverse transcription to cDNA and labelling
Scanning and image processing
Normalisation and pre-processing

Fig. 1. Principal steps of a microarray experiment.

labeled by a dye and hybridised to the microarray probes. The cDNA binds only to specific oligonucleotides on the array because of the base-pair complementarity. After hybridisation, the dye is excited by a laser, so that the amount of annealed cDNA can be quantified by measuring the fluorescence intensities. Using image processing software, the fluorescence background is subtracted and the expression values for each monitored gene are calculated. This process is repeated for every sample. Finally, the data of all samples are incorporated into one table constructing the gene expression matrix $\boldsymbol{G}$. This gene expression matrix is usually the input to a classification system. The rows of this matrix correspond to the single genes and the columns to the single samples. In the following, we refer to the space spanned by the genes as gene space $\mathcal{G}$ of the dimension $d_{\mathcal{G}}$ (with $d_{\mathcal{G}}$ being the total number of genes). The samples can then be considered as vectors in this space $\mathcal{G}$.

We have analysed two publicly available microarray data sets as case studies. The first one is a data set of 72 classification examples for leukaemia, that consists of two classes [8]. The two types of leukaemia are acute myeloid leukaemia (AML) and acute lymphoblastic leukaemia (ALL). The expression values of over 6000 genes and ESTs were monitored by Affymetrix arrays. ALL can be subdivided further into T-cell and B-cell lineage classes. Golub et al. split the data set into 38 cases (27 ALL, 11 AML) for learning and 34 cases (20 ALL, 14 AML) for testing the classifier model [8]. These two sets came from different laboratories. The test set shows a higher heterogeneity with regard of tissue and age of patients. The second data set is composed of gene expression data from 40 tumour and 22 normal colon tissue samples [2]. It was also generated using Affymetrix oligonucleotide arrays. A difficulty in the analysis of this data is the large variation in the tissue composition of the samples. Cancer samples show a strong bias towards epithelial cells, while normal tissue samples contain a mixture of different cell types. This means that differentially expressed genes might not be involved in the cancer development, but are rather specific for the tissue type of the sample. We will see that the heterogeneity of the tissue samples interferes with the performance of the applied classifiers and might pose a considerable challenge to overcome for clinical applications of microarray techniques.

3. Classification of gene expression data—challenges and a methodology

3.1. Challenges in the analysis of microarray data

Classification of tissue samples based on microarray experiments faces several major challenges: (i) Microarray data contain a high level of noise due to experimental procedures [18]. The expression values of single genes are very variable within tissue samples from the same class. Even without experimental noise, genes generally show a large biological variance [9]. (ii) Microarray data are sparse: They usually contain a large number of genes (features), but only a small number of samples [8,2]. (iii) Many genes are highly correlated, which leads to redundancy in the data [7]. (iv) Tissue samples themselves may be wrongly classified by physicians, since results of standard tests can be contradictory or inconclusive [22]. All these reasons make any classification of microarray data difficult. In this section, we discuss the first three of these challenges in microarray data analysis followed by an outline of our methodology to overcome them.

3.1.1. Microarray data inherit large experimental and biological variances

Experimental procedures such as tissue handling, RNA extraction and amplification or hybridisation introduce variances and biases. Some oligonucleotide and cDNA probes may not be specific to one gene and several different genes may hybridise to the same probes (‘cross-hybridisation’). The labeling of cDNA and the scanning of the slides frequently show non-linear behaviour, which makes any correction difficult. With data quality and normalisation methods currently improving, many of the sources of noise might be reduced in the near future. However, as gene expression is highly variable even in different samples of the same tissue, the necessity remains that classifiers for microarray data will have to be robust to a high degree of inherent variance.

3.1.2. Microarray data are sparse

A second major challenge is the high dimensionality of the input space. In present microarray experiments, several thousands of genes are monitored, while the number of samples is often restricted to hundreds or less. It is well known in pattern recognition that classifiers frequently yield poor performance for these kinds of situations, when the number of examples is small compared to the number of features. A large input space usually results in a large number of parameters in the model. Since the number of examples is small, the determination of the model parameters is often insufficient.

3.1.3. Microarray data are highly redundant

Many genes are strongly correlated because of the high connectivity in the genetic network within the cell. Groups of genes often share the same expression pattern if they are included in a specific pathway. These genes are called ‘coexpressed’. Adding coexpressed genes to the classification system does not increase information for the system. Furthermore, many genes on microarrays do not show any change in expression across the experiment. This is due to the fact that microarrays at present are usually not specific in the selected genes that are used as probes. The selection of genes to be included in a microarray experiment is frequently based on their availability and not on their functionality. Microarray data tend, therefore, to have a high degree of redundancy.

3.2. A cancer profiling methodology

In this subsection, we address the issue of feature selection and outline a methodology for profiling a disease (in this case cancer) based on gene expression data. Robust classification and classification of sparse data will be discussed in Section 5.

Feature selection procedures aim at improving the classification by excluding irrelevant features and thereby reducing the size of the input space. Features are excluded if they contribute only weakly to the classification. In our case, a high number of genes do not change their expression between classes. Including these genes introduces noise and may yield poorer classification performance. We wish to select only those genes that can serve as a good basis for discrimination between classes and, thus, for classification. Ideally, we would like to find the intrinsic dimensionality of the data. A d-dimensional data set is said to have the intrinsic dimensionality $\tilde{d}$ if the entire data lies in a $\tilde{d}$-dimensional subspace. The selected genes should show distinct expression patterns in different tissue classes. However, we can expect that different tissue classes have a different intrinsic dimensionality in the gene space. Each class of tissue may have a defined set of genes, the expression levels of which are characteristic for this class. These sets of genes may not be overlapping, e.g. a gene may be highly correlated with single cancer subtype, but not correlated with other subtypes. In selecting an optimal number of genes we have to balance the simplicity with the robustness of the model. Reducing the number of input features leads to a decreased number of model parameters. However, as genes show a high intrinsic variability, the classification should not depend on too few genes.

The incorporation of feature selection in a classification system can be done in two different ways. Feature selection and classification can be treated independently and separately from the classification model. Features are selected with respect to predefined criteria. This approach often has the advantage of being computationally inexpensive and easy to process. In another approach, the selection of features is determined by the classification model itself, since an optimal set of features depends on the choice of the classifier. This constitutes an integrated approach. Unfortunately, the large size of the gene space $\mathcal{G}$ prohibits an exhaustive search for optimal sets of genes if complex classifiers like ANNs are used. Sequential search techniques may still be practical but frequently find only local optima. Additionally, they are rather sensitive to noise.

In this study, we sought to combine the advantages of a separated and an integrated feature selection. We first used filtering and correlation analysis to find a set of single classifying genes (i.e. genes which are independently good features for classification). This set was used for training and testing of our classifiers. To select more complex features, like groups of genes that define a profile of a disease, we applied rule extraction. Finding complex expression profiles may accommodate the complexity of the underlying genetic networks better and enable us to reveal the fine structures of cancer classes. The extracted rules point to groups of genes that are more indicative for a particular tissue type than single marker genes.

The outline of the complete methodology that we introduce in this study are as follows:

(1) *Normalisation*: The scaling of the intensities enables the comparison of expression values between different microarrays within an experiment.
(2) *Preprocessing*: Filtering aims at eliminating the low expressed genes or constantly expressed genes across the tissue classes. The log-transformation helps to balance the range of the data.
(3) *Feature selection*: A set of significantly differentially expressed genes across the classes is selected for clustering and classification.

(4) *Clustering or unsupervised classification*: Grouping of tissue samples with similar expression patterns reveals preliminary profiles of cancer samples.
(5) *Supervised classification*: The tissue classes are modeled by the use of classifiers. The model parameters are optimised and the resulting structure of the model analysed.
(6) *Knowledge discovery*: The extraction of rules refines the profiles for each class and results in complex expression signatures (profiles) of groups of genes. The rules represent the fine grades of the common expression levels within these groups. By using thresholds, smaller or larger groups can be investigated [12].

The details of every step are described in Section 6.

4. Knowledge representation in KBN

Several different methods have been used so far for the classification of cancer tissue. Golub et al. applied a test similar to the *t*-test to find a set of 50 genes to distinguish acute myeloid leukaemia from acute lymphoblastic leukaemia [8]. Based on this set, a classification was achieved by weighted voting. Very recently, ANNs were used to discriminate small round blue cell tumours which present considerable diagnostic problems [13]. Correct diagnosis of these tumours is crucial given that clinical decisions are required regarding selection of drugs, some of which can have dangerous side effects. From within a total set of over 6000 genes, a smaller set was selected by ranking the sensitivity of multilayer perceptrons (MLPs) for features (genes) and used successfully for the class prediction. The disadvantage of MLPs is their 'black box' character, since MLPs offer little or no insight into their decision processes. Knowledge in conventional ANNs like MLPs is stored locally in the connection weights and distributed over the whole network, complicating its interpretation.

Although these approaches could successfully classify tumours and assign single genes a value for their importance within the classification, a major pitfall is the basic assumption of genes as independent variables. Genes are highly correlated and most of the functions within a cell are carried out by a complex interaction of several genes. This biological fact should be reflected in the structure of a classifier. One possibility may be the construction of more complex statistical models but this would be rather difficult since relatively little is known about the complex genetic networks in cells.

To overcome these drawbacks, we applied knowledge-based neurocomputation for the classification of cancer tissues using microarray data. KBN addresses the problem of knowledge representation and extraction [4]. It strives for explicit modeling and harvesting of knowledge in neural networks. This may be done by encoding of prior information as well as refining and extracting knowledge that has been acquired by a neurocomputational system. Neural networks that follow these concepts are called knowledge-based neural networks (KBNN). KBNN have been applied successfully in the field of bioinformatics to classify DNA and RNA sequences [21]. Towell and Shavlik introduced the KBANN algorithm to encode and refine expert knowledge. While this is practical in sequence analysis where knowledge has been accumulated over the past decades and cause–action dependencies have been well studied, this is not the case for gene regulation on the genomic scale. The study of the complex interactions and the various functions of genes is still in the nascent stages. The main goal of a KBNN in microarray data analysis is to (1) model non-linear associations between many genes and the studied diseases, and to (2) extract knowledge, possibly in the form of rules.

The mapping of the structure of a neural network or its input–output behaviour to a set of inference rules is commonly referred as rule extraction. Two main categories of rule extraction exist: decompositional and pedagogical approaches. Decompositional rule extraction techniques derive rules by the study of the structure of the neural network, translating the weights and biases of single units into inference rules. An example for a decompositional approach is ‘structural learning with forgetting’ introduced by Ishikawa [10]. In pedagogical rule extraction, the input–output behaviour of the whole network is approximated by rules. An example for a pedagogical technique is the TREPAN algorithm that seeks to express the functional behaviour of a neural network in a decision tree structure [5].

Another characteristics of KBNNs is the type of rules that a KBNN deal with. We list some of the types of rules that have been represented and extracted from KBNNs [4,11]:

(1) Simple propositional rules (e.g. IF x_1 is A AND/OR x_2 is B THEN y is C, where A, B and C are constants, variables, or symbols of true/false type);
(2) Propositional rules with certainty factors (e.g. IF x_1 is A (CF1) AND x_2 is B (CF2) THEN y is C (CFc), where CF*i* are certainty factors);
(3) Zadeh–Mamdani fuzzy rules (e.g. IF x_1 is A AND x_2 is B THEN y is C, where A, B and C are fuzzy values represented by their membership functions);
(4) Takagi–Sugeno fuzzy rules (e.g. IF x_1 is A AND x_2 is B THEN $y = ax_1 + bx_2 + c$, where A and B are fuzzy values and a, b and c are constants);
(5) Fuzzy rules of type (3) with degrees of importance and certainty degrees (e.g. IF x_1 is A (DI1) AND x_2 is B (DI2) THEN y is C (CFc), where DI1 and DI2 represent the importance of each of the condition elements for the rule output, and the CFc represents the strength of this rule).
(6) Fuzzy rules that represent associations of clusters of data in the input space (e.g. Rule j: IF [$\boldsymbol{x}$ is in the input cluster $\boldsymbol{c}_{j,\mathrm{in}}$ defined by its center (x_1 is A_j, to a membership degree of MD_{1j}, AND x_2 is B_j, to a membership degree of MD_{2j}) and by its radius $R_{j,\mathrm{in}}$] THEN [$\boldsymbol{y}$ is in the output cluster $\boldsymbol{c}_{j,\mathrm{out}}$ defined by its center (y is C, to a membership degree of MDc) and by its radius $R_{j,\mathrm{out}}$, with $N_{\mathrm{ex}(j)}$ examples represented by this rule]).

As a particular version of a KBNN we applied evolving fuzzy neural networks (EFuNNs), that are implementations of evolving connectionist systems (ECOS). EFuNNs deal with rules of type (6) [12]. We describe and justify their use in the next section.

5. Evolving connectionist systems for adaptive learning and rule extraction

ECOS are systems that evolve in time through interaction with the environment, i.e. an ECOS adjusts its structure with a reference to the environment [12]. ECOS are multi-level, multi-modular structures where many modules have inter- and intra-connections.

The evolving connectionist systems are not restricted to a clear multi-layer structure. It has a modular open structure. The functioning of the ECOS is based on the following general principles:

(1) ECOS learn fast from a large amount of data through one-pass training.
(2) ECOS adapt in an on-line mode where new data is incrementally accommodated.
(3) ECOS have an 'open structure' where new features (relevant to the task) can be introduced at any stage of the system's operation, e.g. the system creates 'on the fly' new inputs, new outputs, new modules and connections.
(4) ECOS memorise exemplars (prototypes) for a further refinement, or for information retrieval.
(5) ECOS learn and improve through active interaction with other systems and with the environment in a multi-modular, hierarchical fashion.
(6) ECOS adequately represent space and time in their different scales; have parameters that represent short-term and long-term memory, age, forgetting, etc.
(7) ECOS deal with knowledge in its different forms (e.g. rules, probabilities); analyse themselves in terms of behaviour, global error and success; 'explain' what the system has learned and what it 'knows' about the problem it is trained to solve; make decisions for further improvement.

One particular implementation of ECOS is the evolving fuzzy neural network EFuNN system [12]. A simplified structure of an EFuNN is shown in Fig. 2.

We used fuzzy neural networks to study gene expression because of the following reasons:

- Microarray data have a large noise component. Fuzzy classifiers have shown that they are robust for this kind of data [11].
- The variability of the data is large. Binning gene expression values into categories like 'high' or 'low' in a crisp fashion may lead to over-simplification. The use of fuzzy membership functions is more adequate.
- As still little is known about gene expression on the genomic scale, defining crisp concepts is difficult. Fuzzy variables may be useful.

Fuzzy logic is less sensitive to imprecise data than classical Boolean logic, since in fuzzy logic it is not necessary to set hard thresholds between concepts like 'gene *i* has low expression' and 'gene *i* has high expression'. A certain gene expression value can be partially labelled as 'high' or 'low'. This labelling is defined by a set of membership

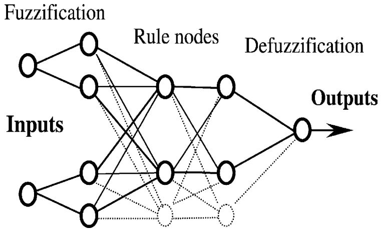

Fig. 2. Outline of a simplified EFuNN structure.

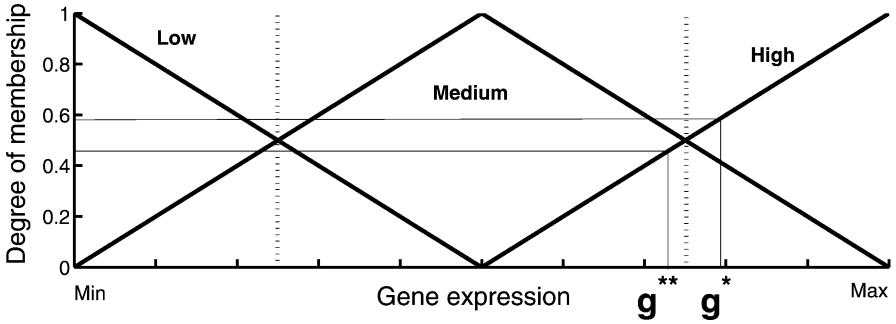

Fig. 3. By using triangular membership functions a crisp gene expression value $g^{\star}$ can be mapped into fuzzy labels. In the above example, $g^{\star}$ has a degree of membership of 0.58 for 'highly expressed' and 0.42 for 'medium expressed'. Classical Boolean set theory set crisp thresholds (dashed lines) for membership. It makes it possible for two similar expressed genes to be put in different categories (e.g. $g^{\star}$ would be labelled as 'highly expressed' and $g^{\star\star}$ as 'medium expressed') making Boolean logic more sensitive to noise.

functions μ_i that represent a mapping of crisp values to fuzzy labels (e.g. 'low', 'medium', 'high'). Fig. 3 illustrates this concept.

EFuNNs incorporate fuzzy labels by the use of a five-layer structure where nodes and connections are created/connected as data examples are presented (see Fig. 2). The first layer serves for fuzzification of the input, so that the activation values in the second layer are the fuzzy representation of the input. The third layer of neurons (rule nodes) evolves through either supervised or unsupervised learning. These rule nodes link hyperspheres in the fuzzy input space to hyperspheres in the fuzzy output space. Examples will be classified according to their distance to the nearest rule nodes. The fourth layer of neurons performs defuzzification, so that the fifth layer represents the real values for the output variables.

Different learning, adaptation and optimisation strategies and algorithms can be applied to an EFuNN structure [12]. Some of these are as follows: (a) Active learning—learning is performed when a stimulus (input pattern) is presented and kept active; this is the main learning mode. (b) Passive (inner, ECO) learning mode -learning is performed when there is no input pattern presented to the EFuNN. In this case the process of further elaboration of the connections in an EFuNN is done in a passive learning phase, when existing connections, that store previously fed input patterns, are used as 'echo' (here denoted as ECO) to reiterate the learning process. Different structure optimisation techniques can be applied during the learning process: (a) Pruning and forgetting -the nodes and connections that are not actively participating in the learning process get pruned according to set criteria; (b) Aggregation and abstraction—rule nodes (each of them representing a centre of a cluster) that are close in the domain space, i.e. accommodate similar exemplars, get merged together into a single new rule node placed in the centre of the aggregated clusters.

In the implementation we applied in this study, the connections between the second layer and the third (rule node) layer evolve by unsupervised learning while supervised learning is used for the connections between the third and the fourth layer. The learning algorithm is described in a shortened version in Fig. 4. Details can be found in [12]. We point out that the rule nodes in EFuNNs form a local representation of the domain space. This contrasts

```
EFuNN training algorithm:
For every fuzzified example $(\mathbf{x_i}, \mathbf{y_i})$ consisting of input value $\mathbf{x_i}$ and target value $\mathbf{y_i}$
    - Find the normalised fuzzy distance $D_j = \frac{||\mathbf{x_i}-\mathbf{r_j}||}{||\mathbf{x_i}+\mathbf{r_j}||}$ to the nodes $\mathbf{r_j}$
    - Calculate the activation $A_j = 1 - D_j$ of all rule nodes $r_j$
    - Define the closest rule node $r_j^\star$ with the highest activation $A_j^\star$
      (for 1-of-N mode)
    - If $A_j^\star < S$ (S: Maximum sensitivity threshold)
        <create a new rule node > with $\mathbf{r_{new}} = \{\omega^{\mathbf{1}}_{\mathbf{r_{new}}} = \mathbf{x_i}, \omega^{\mathbf{2}}_{\mathbf{r_{new}}} = \mathbf{y_i}\}$,
          where $\omega^{\mathbf{1}}_{\mathbf{r}}$: weights of rule node $r$ connecting layers 2 and 3;
                $\omega^{\mathbf{2}}_{\mathbf{r}}$: weights of rule node $r$ connecting layers 3 and 4.
    - Else
        - Propagate the activation of $\mathbf{r_j^\star}$ to the output layer
        - Calculate the error in the output layer $E_{out}(\mathbf{x_i}) = \frac{||\mathbf{y_i}-\mathbf{o_i}||}{n_{out}}$
          ($\mathbf{o_i}$: activation of the output layer for the example $i$; $n_{out}$: number of
          outputs)
        - If $E_{out} > Errth$ (Error threshold)
            <create a new rule node> with $\mathbf{r_{new}} = \{\omega^{\mathbf{1}}_{\mathbf{r_{new}}} = \mathbf{x_i}, \omega^{\mathbf{2}}_{\mathbf{r_{new}}} = \mathbf{y_i}\}$
        - Else update the connection weigths between layers 2 and 3 as well as
          3 and 4. ($\alpha$: learning rate):
          $\omega^{\mathbf{1}}_{\mathbf{r^\star}} = \omega^{\mathbf{1}}_{\mathbf{r^\star}} + \alpha(\mathbf{x_i} - \omega_{\mathbf{r^\star}})$
          $\omega^{\mathbf{2}}_{\mathbf{r^\star}} = \omega^{\mathbf{2}}_{\mathbf{r^\star}} + \alpha\, E_{out}\, A_j^\star$
    - If $mod(i, nAgg) = 0$ (Aggregation after $nAgg$ training examples)
        < aggregate rule nodes, based on aggregation criteria >
```

Fig. 4. EFuNN training algorithm (shortened version; for details see [12]).

to MLPs with a distributed representation of the input examples by the activation values in the hidden layer. In an MLP, the hidden nodes with sigmoid activation functions span hyperplanes in the domain space. A new example is classified based on its position with respect to these hyperplanes. This is problematic if the data are sparse, as it is the case of microarray data. Examples should not be classified at all if they have very different expression profiles than the ones in the training set to avoid crucial misclassifications. This can be easily achieved by a classifier with a local representation in the domain space by setting appropriate activation thresholds. New examples with novel gene expression profiles stay below the threshold and remain unclassified.

The knowledge discovery within the EFuNN framework is done by extracting fuzzy rules (see Fig. 5). Since the knowledge is gained by analysing the internal structure of the neural networks, this approach belongs to the decompositional extraction methods in KBN. Rule extraction is fairly straightforward due to the one-to-one correspondence between rule nodes and produced rules. In contrast to MLPs, the learned knowledge is locally embedded and not distributed over the whole neural network. The granularity of these rules is defined by setting a maximum receptive field of a rule node. Decreasing the size of the maximum receptive field leads to an increased number of rule nodes as the aggregation is restricted.

Rule extraction algorithm: The membership degree I^i_{jk} of a fuzzy label f^i_j of the feature j in a rule k is

$$I^i_{jk} = \frac{\partial \vec{\mathbf{r}}_\mathbf{k}}{\partial f^i_j} = \omega^{1i}_{jk}$$

In the same way a membership degree C^l_k of a rule k for the class l is given by

$$C^l_k = \frac{\partial O^l_k}{\partial A_k} = \omega^{2l}_k$$

where O^l_k is the output of the rule node k to the class node l and A_k is the activation of rule node k. Using this and the EFuNN structure based on fuzzy labels, every rule node r_k will be mapped to a rule R_k:

$$(\mathbf{r_k}, \omega^{\mathbf{1}}_{\mathbf{k}}, \omega^{\mathbf{2}}_{\mathbf{k}}) \rightarrow (R_k, \mathbf{I_k}, \mathbf{C_k})$$

Example rule:
IF Gene X is low expressed (0.9) AND Gene Y is highly expressed (0.8)
THEN Sample i is cancer tissue (0.95).
(Radius of the spherical cluster - 0.3; number of examples covered by this rule - 15).

Fig. 5. EFuNN rule extraction algorithm (shortened version; for details see [12]).

In the extreme case, every created rule node corresponds to just one example. EFuNNs can then be seen as a kind of fuzzy k-nearest neighbour classifier.

Although different membership functions can be exploited within EFuNNs, the usage of the triangular membership function guarantees the conservation of the property $\sum_i \mu_i(g_{jk}) = 1$, where μ_i is the ith membership value for the gene expression value g_{jk} of gene j in sample k. This results from the specific form of the supervised learning method which updates the rule node vector r_j according to $\tilde{r}^i_j = r^i_j + \alpha(x^i_k - r^i_j)$, ($x^i_k, r^i_j$: the membership degrees of the fuzzy variable i for the kth example and rule node j; α: learning rate). It follows that

$$\sum_i \tilde{r}^i_j = \sum_i (r^i_j + \alpha(x^i_k - r^i_j)) = \sum_i r^i_j + \alpha \sum_i x^i_k - \alpha \sum_i r^i_j = 1 - \alpha + \alpha = 1$$

This property ensures that the rules extracted from an EFuNN are consistent in the anterior part. Contradictory rules of the form [IF g_i is *high* ($\mu^{\text{high}}_i = 1$) AND g_i is *low* ($\mu^{\text{low}}_i = 1$) THEN sample k is cancer tissue] cannot be created.

6. Modeling and knowledge discovery from gene expression data on the case studies of cancer tissues

In this section, we apply the described methodology to the leukaemia and colon cancer data sets [8,2]. We first normalise and preprocess the data and examine the importance of these steps for the classification. Using EFuNNs, we classify the data sets and analyse the results and structures of the fuzzy neural networks. Finally, we use rule extraction to identify groups of genes that form profiles and are highly indicative of particular cancer types.

6.1. Normalisation, preprocessing and feature selection

The first step in the analysis of microarray experiments is the normalisation of the data. Subtle differences in hybridisation conditions or sample preparation may lead to large variations in the intensities across different slides. A very simple normalisation procedure is to scale the arrays to have the same total fluorescence [2]. This is based on the assumption that only a small fraction of the genes on the array change their expression, so that the total amount of messenger RNA stays approximately the same in different tissues [16]. Several recently published studies have attempted to improve this rather crude normalisation method. However, no procedure has been established as the most favourable (and it is doubtful that there will be a 'best' method, as we discuss this issue later). For the purpose of this study, we scaled the arrays to have the same total fluorescence intensity.

The next step in the analysis is preprocessing of data. Since microarray data are frequently strongly skewed, using the $\log_2$-transformation helps to balance the data. To exclude genes which do not show changes in expression, we filtered out all genes k with $[\max(\log_2(g_k)) - \min(\log_2(g_k))] < 3$. Although these procedures are common in the analysis of microarray data [19,15], they seemed somewhat arbitrary. A major direction in our present research is, therefore, to find an optimal adjustment of the preprocessing and normalisation within the classification process (see Section 7).

For the selection of genes for classification, several different methods like t-test or Fisher discriminant analysis have been proposed [2,6]. As the main focus of this study was not the selection of genes, we used a rather simple method. We selected genes if they are highly correlated with the tissue classes. Since we do not want to discriminate between correlated and anti-correlated genes, we used the squared Pearson correlation coefficient c_{P}^2 that is computed as follows:

$$c_{\mathrm{P}}^2(\boldsymbol{g}_i, \boldsymbol{c}) = \left[\frac{1}{N}\sum_j \left(\frac{(g_{ij} - \bar{g}_i)(c_j - \bar{c})}{\parallel \boldsymbol{g}_i \parallel \parallel \boldsymbol{c} \parallel}\right)\right]^2$$

with the vector $\boldsymbol{g}_i$ of the expression value of gene i, the class membership c_j of the samples j and $\parallel \cdot \parallel$ the Euclidean distance. Although it is a simple approach for gene selection, it finds sets of genes that are well suited for discrimination of the tissue classes. This can be seen through using principal component analysis (PCA), which gives a first insight in the structure of high-dimensional data. PCA consists of a linear transformation from the original set of variables to a new (smaller) set of orthogonal variables (principal components) so that the variance of the data is maximal and ordered according to the principal components. After selecting 100 genes, the visualisation of the first two components of the leukaemia data set shows that AML and ALL samples are generally well separated with only a small overlap between the two classes (see Fig. 6). For the second data set, the colon cancer and normal tissue samples are less distinguished and include several outliers (see Fig. 9).

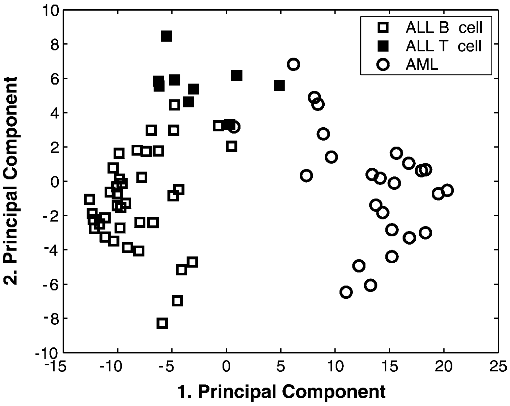

Fig. 6. PCA of leukaemia samples based on 100 genes that have the largest squared correlation with the two classes ALL and AML. The first two principal components include 63.3% of the total variance of the data.

Besides distinguishing between the tissue classes, selected genes should not depend on a specific choice of the data samples. To examine the stability of selection, we used a bootstrapping method. Leaving out one sample, the genes are selected based on the remaining $(N - 1)$ samples. This is then repeated for all samples. Genes that are well correlated across all samples should be frequently selected, while genes that are correlated only with a subset of the samples should be selected less often. Fig. 7 shows most of the genes are chosen repeatedly. This demonstrates that the selection using squared Pearson correlation is highly stable at least for the data sets analysed here.

We finally assessed the quality of normalisation and preprocessing by testing their influence on the classification performance using N-fold cross-validation. (Briefly, the N-fold cross-validation procedure consists of withholding one sample for testing while the classifier is trained with the remaining samples. After repeating this for all samples, the

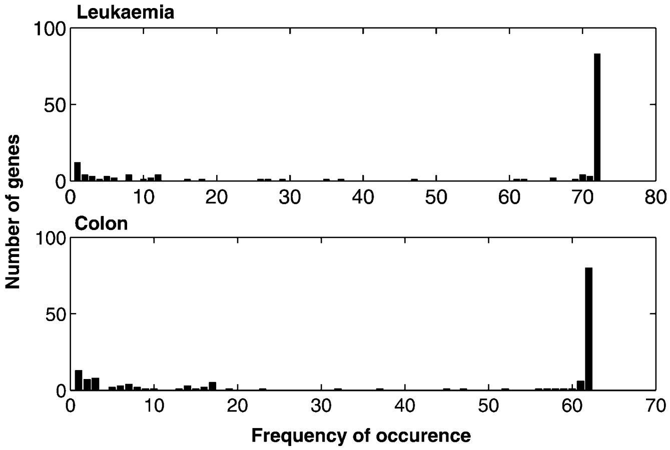

Fig. 7. Stability of gene selection by squared Pearson correlation for the leukaemia and colon data set. The graphs show how frequently genes were selected applying the ‘leave-one-out’ method.

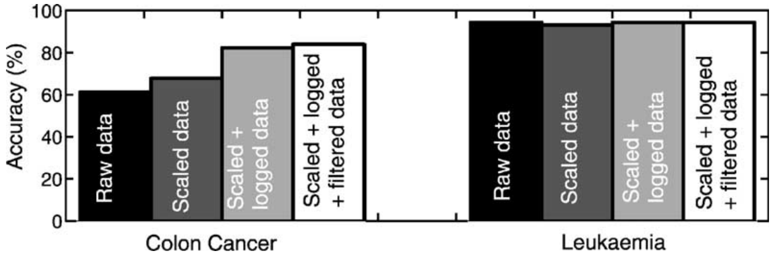

Fig. 8. Dependency of performance by EFuNN on preprocessing and normalisation of data. The *N*-fold cross-validation testing was based on 100 selected genes (EFuNN parameter: error threshold = 0.9; maximum receptive field = 0.3; aggregation number = 20).

cumulative performance is averaged.) As it can be seen in Fig. 8, normalisation and preprocessing can strongly influence the classification performance, as in the case of colon tissue data. The performance was increased by normalisation and further by log-transformation and filtering of the data. For leukaemia, however, the improvements were minor. Two explanations are possible: either the leukaemia data contained less variation than the colon cancer data due to experimental procedures, or our specific choice of normalisation and preprocessing steps (in conjunction with the gene selection) was better suited for colon cancer than for leukaemia. This suggests that normalisation, preprocessing and feature selection are not only strongly linked to the classification performance but are also data dependent. A future classification system may adjust these pre-classification steps to the data to improve the performance (see Section 7).

6.2. Classification

The performance of EFuNNs was tested by *N*-fold cross-validation. The cross-validation procedure included gene selection to avoid any bias due to the selected genes. For modeling, the parameters of EFuNNs have to be adjusted. Note that we used three triangular membership functions 'low', 'medium' and 'high' throughout the analysis (see Fig. 3). The chosen parameters not only determine the classification accuracy but also the internal structure of EFuNNs by controlling the number of rule nodes. Frequent aggregation of rule nodes and a larger maximum receptive field, for example, yield a smaller number of rule nodes. An example of this behaviour is given in Table 1 for the leukaemia data.

Table 1
Performance of EFuNNs in *N*-fold cross-validation testing

Data/classifier	Errth	maxRF	nAgg	Rules	Acc-tr	Acc-te
Leukaemia/EFuNN	0.9	0.3	20	6.1	97.4	95.8
Leukaemia/EFuNN	0.9	0.5	20	2.0	95.2	97.2
Colon/EFuNN	0.9	1.0	40	2.3	88.8	90.3
Colon/EFuNN (compacted)	0.9	1.0	40	2.3	88.8	91.9

Errth: error threshold; maxRF: maximum receptive field; nAgg: aggregation number; Rules: average number of evolved rule nodes; Acc-tr: average classification accuracy of training examples (%); Acc-te: average classification accuracy of testing examples (%). (For a detailed description of the EFuNN parameters see [12].)

By reducing the maximum receptive field from 0.5 to 0.3, the average number of produced rule nodes during the training phase increased from 2.0 to 6.1. While the classification accuracy rose for the training set, it decreased for the testing set. This is a consequence of the well-known bias-variance trade-off. Classifiers of high flexibility tend to overfit resulting in a poorer generalisation. They achieve a higher accuracy for the training data, but show a reduced performance on the test data. For the leukaemia data set, a maximum accuracy by EFuNNs of 97.2% was obtained on the testing set using *N*-fold cross-validation. This is a very high accuracy considering that the trained EFuNN models consisted in average of only two rules. Golub et al. achieved a lower performance (94.3%) applying a weighted voting method [8]. However, a direct comparison is restricted since Golub et al. assessed the accuracy based on a single split of the total data in training and testing set which may lead to biased estimates.

The classification performance is considerably lower for the colon cancer data with 90.3%. (We will improve the accuracy by EFuNN rule compaction in the next section.) Fisher's linear discriminant analysis conducted by Xiong et al. achieved an accuracy of 87.0% [23].

For the colon data, the PCA shows several samples as outliers (see Fig. 9). This made the classification more difficult. As for the leukaemia data set, a large maximum field ensured the simplicity of the model and a high accuracy on the test set. The misclassifications of samples are shown as a PCA projection in Fig. 9. Most of these misclassifications can be considered as outliers. Based on the set of selected genes and the PCA projection, their gene expression profiles were very similar to expression profiles of samples, that belong to the other tissue class. This shows that the selection of genes for classification is crucial to achieve a good separation of the tissue classes.

To elucidate this point further, we compared the muscle index for the samples of the colon cancer data set with the misclassifications by the trained EFuNN. The muscle index was introduced by Alon et al. to obtain a qualitative measure for the muscle content in the

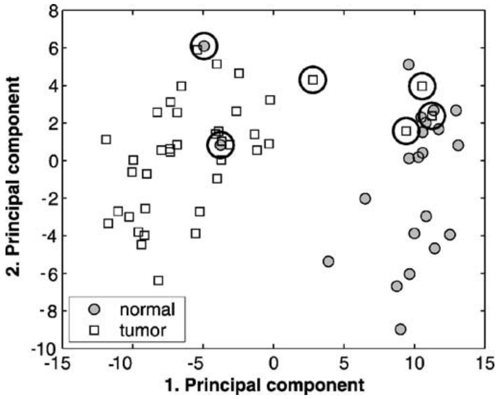

Fig. 9. PCA of colon samples with misclassifications. Classification was based on 100 selected genes (EFuNN parameter: error threshold = 0.9; maximum receptive field = 1.0; aggregation number = 40). Misclassifications are observed when there are very similar gene expressions of two tissue samples, each tagged with a different class label.

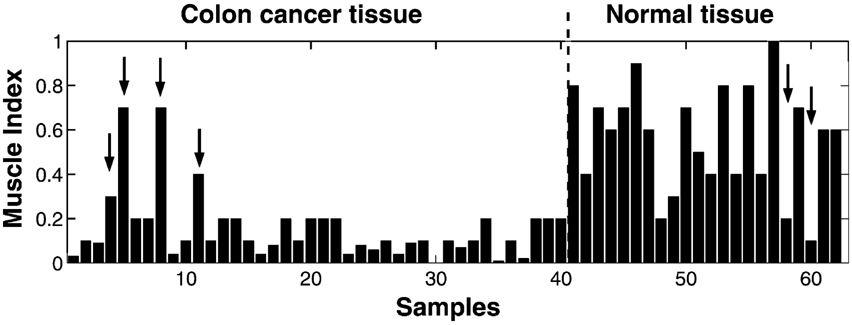

Fig. 10. Muscle index of the samples in the colon tumour data set as defined in [2]. Arrows indicate the misclassified samples by EFuNN. Samples 1–40 are cancer tissue samples whereas samples 41–62 are normal tissue samples. The muscle index is normalised to vary between 0 and 1.

studied samples [2]. It is calculated by averaging over the expression values of 17 ESTs that are homologous to smooth muscle genes. Tumour samples have generally a low muscle index, as they are biased towards epithelial tissue. Normal tissue samples have a high muscle index (Fig. 10). The examination of the muscle indices of the misclassified samples shows that the misclassified tumour samples have the highest muscle indices of all tumour samples whereas the misclassified normal tissue samples have the lowest muscle indices of all normal tissue samples. The values of the muscle index for the misclassified samples are similar to the values for the other tissue class. This means that in a sense the trained EFuNN might have correctly predicted the tissue classes based on the underlying dominant types of tissue in the samples, which differ, however, from the labelled classes (tumour/normal) leading to misclassifications. This result demonstrates the importance of tissue purity for a correct classification based on microarray data. It also points to a potential obstacle for microarray techniques to be used in the clinical application. Biopsies of cancer usually contain, besides tumour tissue, various amounts of different kinds of normal tissue. Since each type of normal tissue has its own specific expression profile, the gene expression values measured by microarrays derive from a mixture of expression values determined by the tissue composition of the biopsy. Large variation in the tissue composition of the samples can cloud the separation that we are interested in; namely the separation of tissue samples in tumour and normal. Misclassifications of samples might occur if the tissue composition in a new sample differs strongly from that of the samples on which the classification system is trained. An alternative to the extraction of RNA from heterogenous samples is the so called laser capture microdissection (LCM), which uses laser beams to isolate homogenous tissue. LCM, however, requires a high level of resources and has mainly been restricted to medical research so far. For clinical practise, it is, therefore, important that a classification system can cope with heterogenous tissue samples.

A important question is how many genes are needed for optimal classification. Computationally, this is a very challenging problem, since the model parameters can be expected to depend on the dimensionality of the input. Optimisation of the model includes both the choice of optimal parameters for the classifier and the optimal number

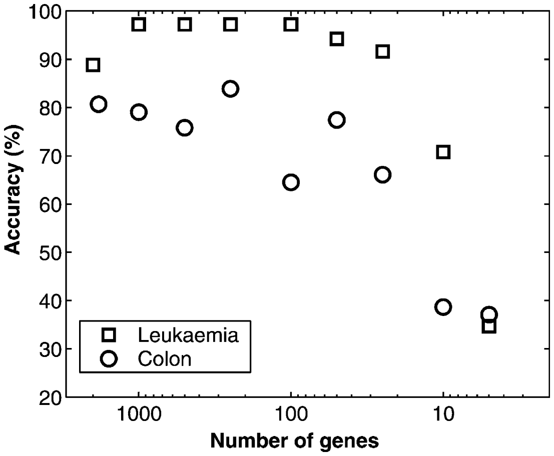

Fig. 11. Dependency of accuracy on the test set using N-fold cross-validation on number of selected genes (EFuNN parameter: error threshold = 0.9; maximum receptive field = 0.9; aggregation number = 20). Too few (or too many genes) used in the classification lead to decreased performance.

of genes. In this study, we restricted ourselves to examining the dependency of the performance on the number of genes as inputs for a fixed set of parameters. A general tendency can be seen in Fig. 11. The classification performance decreases if the number of selected genes is reduced below a certain threshold. Adding more genes to the classification process yields higher accuracy for the leukaemia data set until an optimal performance is reached. A further increase in the number of included genes results in a slight decrease in accuracy. The same general behaviour was observed for the colon tissue data, however, the appearance of several local maxima reflects the higher sensitivity on the particular number of selected genes. Together with the results in Figs. 8 and 9, this indicates that the colon data set has a more complex structure and special care needs to be taken in the preprocessing and feature selection steps.

Before we turned to the extraction of rules, we sought some insight in the internal structure of EFuNNs that have been applied to classification of microarray data. To this purpose, we selected 50 genes by using the squared Pearson correlation. Based on this set of genes, all samples were used for the training and the testing of EFuNNs. The parameters were adjusted to avoid any misclassifications while the number of rule nodes should be as low as possible. This restriction ensures the simplicity of the model and avoids overfitting. The results are presented in Fig. 12. The right side shows the expression matrix and the left side the two main principle components of the data. Two main clusters are apparent, and include most of the samples (43 out of 62). Clearly, the samples that were assigned to a rule node have a very similar expression profile. Outliers shown in the PCA projection on the left side of Fig. 12 have been placed into singular nodes, i.e. nodes including only one sample.

Since EFuNNs use a local representation of the domain space and fuzzification, they offer the possibility of interpreting each rule node separately in a meaningful way. To demonstrate this, we extracted the weights between the second and third layer for a rule

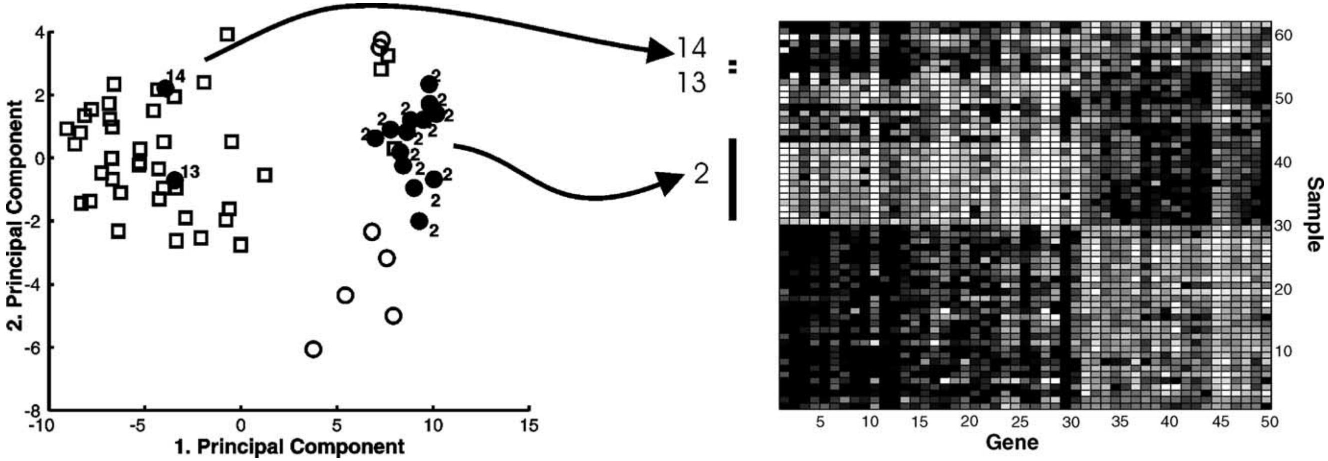

Fig. 12. PCA (left side) and expression matrix (right side) for colon tissue samples. The genes in the expression matrix are sorted according to their correlation with the tissue classes. The samples are sorted according to the labeling by the rule nodes in the trained EFuNN classifier. High normalised expression values are presented as light and low expression values as dark squares in the expression matrix. The numbers in the PCA figure refer to the labels of the rule nodes. Rule node 2 comprises many examples and is associated with a major cluster of normal tissue samples, while rule nodes 13 and 14 represent outliers.

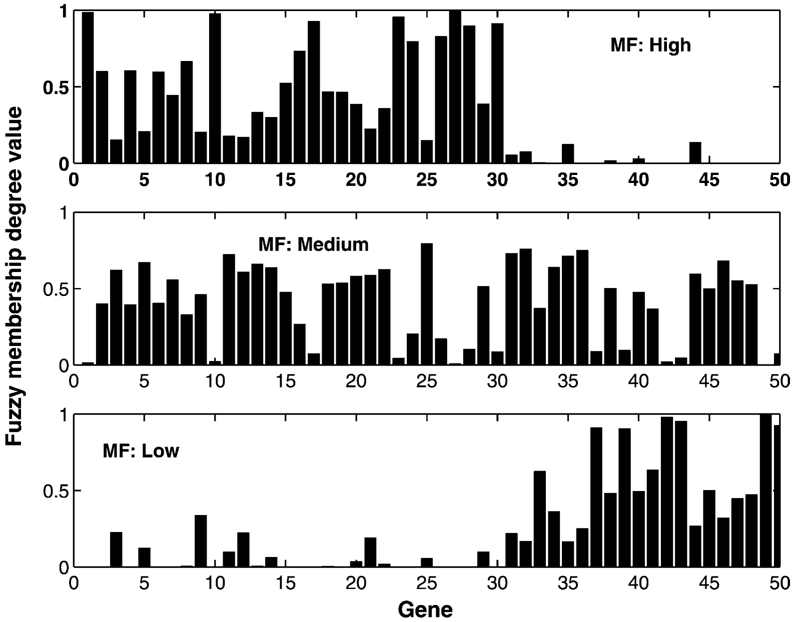

Fig. 13. Distribution of fuzzy membership degrees values within a node. The corresponding values of the three fuzzy labels 'low, medium, high' are shown for rule node 2 in Fig. 12. The membership values for each gene sum up to 1.

node obtained after the training (see Fig. 13). They present the fuzzy membership degree of each fuzzy label for classification by the rule nodes as defined in Fig. 5. If, e.g. a high degree is found for the fuzzy label 'low' for the expression of a certain gene, then a low expression value of this gene in a particular sample favours this sample being assigned to the rule node. Thus, Fig. 13 indicates which gene contributes at which expression level to the classification by a rule node. An interesting observation can be made if we compare the distribution of membership degrees of fuzzy labels with the variance of the expression values within the rule node. Genes with a low variation in expression across the samples achieved a high membership degree for a fuzzy label, e.g. gene 10 for fuzzy label 'high' in Figs. 13 and 14. This is contrasted by genes (like gene 20 in Figs. 13 and 14) with a large

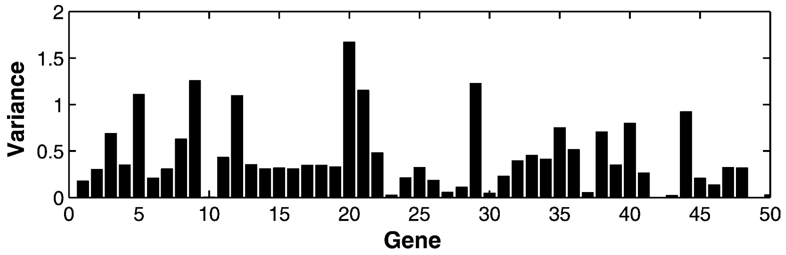

Fig. 14. Variance of gene expression across the samples represented by rule node 2 in Fig. 12. The higher the variance of a gene's expression values, the lower its significance is in the corresponding rule.

variance. Their fuzzy membership is distributed over different fuzzy labels. (Note that the membership degrees for a gene over all fuzzy labels add up to one.) EFuNNs assigned, therefore, the fuzzy membership degree of genes according to their stability in expression. Simultaneously, they selected a set of genes that is indicative for the group of samples belonging to a rule node. This leads us now to the final part of the experiments where we sought for the discovery of knowledge from the trained networks.

6.3. Knowledge discovery through rule extraction

The EFuNN rule extraction algorithm represents the internal structure of the trained network as a set of fuzzy rules. Every rule node obtained during the training phase represents a rule. The rules are 'local' and each of them corresponds to the dominating gene expression within a particular cluster of samples. The rules present the knowledge in a comprehensible form that is accumulated in the network.

To achieve compact rules, we applied a threshold to the fuzzy membership degree of genes and neglected all genes that fall below this threshold. This procedure is motivated by the observation in the last section that a high degree of fuzzy membership indicates a low variability in the expression of a particular gene within the cluster defined by a rule node. By applying a threshold to the fuzzy membership values we select genes that show a stable expression. Thereby we introduced, for each rule node, its own metric in the gene space $\mathcal{G}$. The threshold for the membership degree of genes was stepwise increased until the classification performance drops. In other words, we 'shaved-off' redundant genes in every rule node, obtaining very compact rules and in some cases an improved performance (see Table 1).

We illustrate this approach by applying it to the colon cancer data set. Fig. 15 shows the dependence on the threshold value of the average number of genes included in the extracted rules. Increasing the threshold leads to a considerable decrease in the number of genes participating in a rule. Does the reduced number of genes cause a sacrifice in the accuracy of the rules compared to the original network? To evaluate this possibility, we recorded the

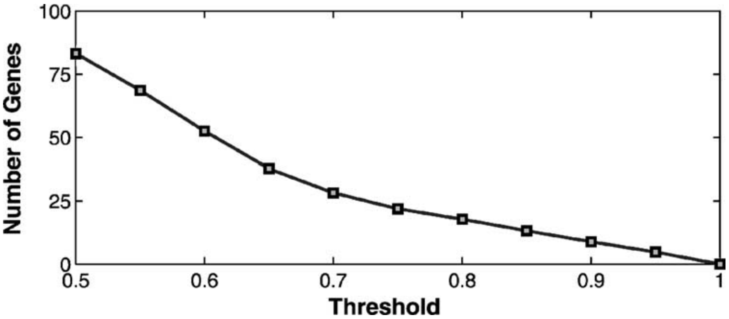

Fig. 15. Average number of genes in the antecedent part of rules extracted from EFuNNs after rule compaction was applied. The average number of genes for the rules is weighted by the number of samples that trigger the rule. The setting of the threshold for the fuzzy membership degree in the rule compaction procedure determines the number of the genes included in the rules.

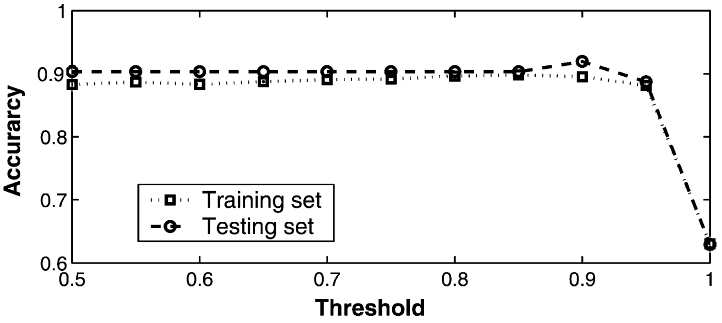

Fig. 16. Overall accuracy of the rules from EFuNNs for colon cancer data in relation to the applied threshold for rule compaction. Both the classification accuracies obtained in an N-fold cross-validation procedure for the training and test samples are presented.

accuracy of the extracted rules on the training and test set. The accuracy of extracted rules on the training set is commonly referred to as the fidelity of the rules. The results are shown in Fig. 16. The performance of the rules remains surprisingly stable over a large range of thresholds. Thus, for an accurate classification only a small number of genes is needed. Since fidelity and accuracy of the extracted rules behave similarly, we can set the threshold to 0.9 without sacrificing the performance of the rules compared to the original network. The average number of genes included in the rule drops to less than 9. A further increase of the threshold would lead to even more compact rules, but also to poorer performance.

An extracted rule is presented in Fig. 17. After the application of the 'rule compaction' algorithm the rule is based on the values of only 9 genes compared to the original 100 genes. This rule is instructive for finding gene markers. Some of the genes found are already known to be linked to colon cancer. *Caveolin* (Z18951) is a putative tumour suppressor gene corresponding to a structural membrane protein involved in the regulation of signaling pathways [17]. Low expression of *caveolin* may facilitate the development of colon cancer. The enzymes carbonic anhydrase I (R93176) and II (J03037) have been shown to be correlated with the aggressiveness of colorectal cancer [3]. MCM3 (H09351) is involved in DNA replication. Its upregulation can be expected since rapid growth of tumour cells results in an increased rate of DNA replication. The underexpression of

Rule for Colon cancer:
IF H57136 is low (1.00) AND H09351 is high (0.92) AND T46924 is low (0.90) AND Z18951 is low (0.97) AND R695523 is low (0.98) AND J03037 is low (0.98) AND R93176 is low (0.97) AND H54425 is low (0.964) AND T55741 is low (0.99)
THEN Sample is cancer tissue (1.0)

Fig. 17. A simplified exemplar rule for colon cancer tissue (EFuNN parameter: error threshold = 0.9, maximum receptive field = 0.1, rule compacting threshold = 0.9, aggregation number = 20). Genes are presented by their GeneBank Accession numbers. The numbers in brackets correspond to the weights in the trained EFuNN. They indicate the importance of the gene in the anterior part of the rule respectively the confidence of the inference for the posterior part of the rule.

smooth muscle myosin light chain kinase (T55741) reflects the fact that normal colon tissue samples were biased towards muscle tissue in the analysed data set [2]. Other genes presented in this rule may serve as candidates for future developments of genetic markers. We stress, however, that the main strength of the rule extraction is not the indication of new single marker genes, but the detection of groups of genes whose cumulative behaviour characterise a particular tissue type. This approach may lead to the identification of new and more complex phenotypes of cancer which cannot be described by the use of single genetic markers only.

7. Conclusions and future directions

Many applications of information systems in the field of bioinformatics dealing with classifying new examples on the basis of previous cases frequently yield very good results. However, with new emerging array technologies and with new DNA data continuously becoming available, research will focus more and more on revealing underlying principles. Besides performance, the ability to discover new knowledge will be a crucial point for every AI system in the field of biological research. Application in the medical field demands further that the classification process is understandable for humans.

Biological and medical research using large-scale gene expression data sets derived from microarray experiments is mainly data driven, starting with little prior information about the underlying genetic network structures. Several different methods for classification of microarray data have been proposed, with the major focus on the classification performance. However, we feel that the discussion about the comprehensibility of these methods and their capacity for knowledge discovery is largely neglected.

In this study, we showed that the criteria for performance and comprehensibility can be met by using KBNNs and in particular EFuNNs as classifiers. The achieved classification accuracy is similar to or better than compared to other standard classification approaches. A major difference between EFuNNs and previously used classifiers is that frequently the latter offer little insight into their decision making process. Applying EFuNNs to leukaemia and colon cancer microarray data, we achieved profiles for the specific cancers and discovered fingerprints of putative cancer subclasses. EFuNNs incorporate unsupervised and supervised learning schemes. This is favourable for knowledge discovery in microarray data. Besides the classification of currently known cancer types, the unsupervised learned structure of EFuNNs might indicate new subtypes of cancer. The discovery of new disease subtypes is an important field of medical microarray analysis. Recent studies have shown that new distinct cancer subclasses found by unsupervised clustering analysis of microarray data have specific clinical characteristics, e.g. the survival rates of patients undergoing chemotherapy [1]. Finding these new disease subclasses leads the way to more disease-specific treatments and a better tailoring of the therapy to the individual patient.

The substructures in the gene expression data can be translated into comprehensible inference rules. Using rule extraction, we found compact rules that are highly indicative for the analysed tissue types. Rules correspond either to clusters of tissue samples or outliers. The rules were defined by the rules nodes in a trained EFuNN. Applying the novel techniques of rule compaction, we derived inference rules that included only a small number of indicative genes without sacrificing the accuracy of the trained network by rule extraction. This approach may give rise to more powerful and flexible medical classification systems of cancer types that overcome the limitations posed by the usage of single marker genes. As in other fields, a total set can be more than just the sum of its objects. The ability to represent the learned knowledge of the neural network in the form of comprehensible rules is also important in respect of the medical application of microarray techniques. The correct classification of tissue samples is crucial for the choice of treatment by the physician. For such safety-critical applications, rules are preferable to the 'black box' approaches.

We elucidated the difficulties for correct classification in dealing with heterogenous tissue samples which is commonly the case in medical practise. To our knowledge, previous studies have neglected the impact of tissue heterogeneity on the performance of a classification system based on microarray data. Alon et al. pointed out the importance of tissue purity in [2], but only in the framework of unsupervised hierarchical clustering. This finding has consequences for possible applications of microarray techniques: First, tissue heterogeneity can interfere with the detection of specific makers for tumour types. For example, using rule extraction we found a muscle gene (which is probably not involved in the cancer development) important for the classification. Purity of the tissue samples is, therefore, of importance for medical research. Second, for clinical applications, classification systems based on microarray data have to be robust against large variations in the tissue composition. We are currently working on methods of correcting for the variation in tissue composition to avoid misclassifications caused by tissue heterogeneity.

To our knowledge, this study is the first showing the possibility of translating microarray-based classifiers into comprehensible sets of rules without sacrificing classification performance. The proposed methodology for gene expression data analysis, modeling and knowledge discovery is a generic one. It can be applied to tissues other than cancer tissues for the discovery of genetic disease profiles. The analysis also shows the importance of a holistic approach towards the tissue classification based on microarray data. Normalisation, preprocessing and feature selection can strongly influence the classification performance. We see the necessity of incorporating the various phases into an enlarged classification system. This integration is a major direction of our current research.

Acknowledgements

We would like to thank the reviewers for their constructive comments and Bronwyn Carlisle for proof-reading. Commercial use of the techniques described and applied in this paper is subject to obtaining a license from Pacific Edge Biotechnology Ltd (http://www.pacificedgebiotech.com).

References

[1] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.

[2] Alon U, Barkai N, Notterman DA, Gish GK, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS 1999;96:6745–50.

[3] Bekku S, Mochizuki H, Yamamoto T, Ueno H, Takayama E, Tadakuma T. Expression of carbonic anhydrase I or II and correlation to clinical aspects of colorectal cancer. Hepatogastroenterology 2000;47:998–1001.

[4] Cloete I, Zurada J, editors. Knowledge-based neurocomputing. Cambridge: MIT Press; 2000.

[5] Craven MW, Shavlik JW. Extracting tree-structured representations of trained networks. In: Touretzky D, Mozer M, Hasselmo M, editors. Advances in neural information processing systems, vol. 8. Cambridge: MIT Press; 1996. p. 24–30.

[6] Dudoit S, Fridlyand J, Speed T. Statistical methods for identifying differentially expressed genes in replicated cDNA experiments. Technical Report 576. Department of Statistics, University of Berkeley; 2000.

[7] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. PNAS 1998;95:14863–8.

[8] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression. Science 1999;286:531–7.

[9] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, et al. Functional discovery via a compendium of expression profiles. Cell 2000;102:109–26.

[10] Ishikawa M. Structural learning with forgetting. Neural Networks 1996;9:509–21.

[11] Kasabov N. Foundations of neural networks, fuzzy systems, and knowledge engineering. Cambridge: MIT Press; 1996.

[12] Kasabov N. Evolving fuzzy neural networks for on-line learning, reasoning and rule extraction. IEEE Trans Syst Man Cybernet Part B Cybernet 2001;31(6):902–18.

[13] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7: 673–9.

[14] Liotta L, Petricoin E. Molecular profiling of human cancer. Nat Rev Genet 2000;1:48–56.

[15] Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumor. Nature 2000;406:747–52.

[16] Quackenbush J. Computational analysis of microarray data. Nat Genet Rev 2001;2:418–27.

[17] Razani B, Schlegel A, Liu J, Lisanti MP. Caveolin-1, a putative tumour suppressor gene. Biochem Soc Trans 2001;29:494–9.

[18] Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, et al. Normalization strategies for cDNA microarrays. NAR 2000;28(10):e47.

[19] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organising maps: methods and applications to hematopoietic differentiation. PNAS 1999;96:2907–12.

[20] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet 1999;22:281–5.

[21] Towell GG, Shavlik JW. Knowledge-based neural networks. Artif Intell 1994;70:119–65.

[22] West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. PNAS 2001;98:11462–7.

[23] Xiong M, Jin L, Li W, Boerwinkle E. Computational methods for gene expression-based tumor classification. Biotechniques 2000;29:1264–70.