Integrated optimisation method for personalised modelling and case studies for medical decision support

Nikola Kasabov* and Yingjie Hu

Knowledge Engineering and Discovery Research Institute,
Auckland University of Technology,
Auckland 1010, New Zealand
E-mail: nkasabov@aut.ac.nz
E-mail: rhu@aut.ac.nz
*Corresponding author

Abstract: Personalised modelling aims to create a unique computational diagnostic or prognostic model for an individual. The paper reports a new Integrated Method for Personalised Modelling (IMPM) that applies global optimisation of variables (features) and neighbourhood of appropriate data samples to create an accurate personalised model for an individual. The proposed IMPM allows for adaptation, monitoring and improvement of an individual's model. Three medical decision support problems are used as illustrations: cancer diagnosis and profiling; risk of disease evaluation based on whole genome SNPs data; chronic disease decision support. The method leads to improved accuracy and unique personalised profiling that could be used for personalised treatment and personalised drug design.

Keywords: personalised modelling; optimisation; data analysis.

Reference to this paper should be made as follows: Kasabov, N. and Hu, Y.J. (2010) 'Integrated optimisation method for personalised modelling and case studies for medical decision support', *Int. J. Functional Informatics and Personalised Medicine*, Vol. 3, No. 3, pp.236–256.

Biographical notes: Nikola Kasabov is the Founder and Director of KEDRI, a Chair of Knowledge Engineering at the School of Computer and Information Sciences at Auckland University of Technology (AUT), a Fellow of IEEE, Fellow of the Royal Society of New Zealand and Fellow of the New Zealand Computer Society. He holds a MSc and PhD from the Technical University of Sofia. His main research interests are in the areas of intelligent information systems, soft computing, neuro-computing, bioinformatics, brain study, speech and image processing, data mining and knowledge discovery.

Yingjie Hu is currently a Postdoctoral Research Fellow in KEDRI at Auckland University of Technology (AUT), New Zealand. He completed his PhD study in Computer Science from AUT in 2011. He has received his MCIS Degree (Hons with 1st Class) in Computer and Information Sciences from AUT in 2006. He received the Top Achiever Doctoral

Scholarship from New Zealand Tertiary Education Committee in 2007. His research interests are in the areas of neural network computing, evolutionary computation, bininformatics and data mining techniques.

1 Introduction

Most contemporary medical decision support systems use global models for the prediction of a patient's risk to develop a particular disease or a likely outcome from the treatment. A global model is derived from all available data for the target and then applied to any new patient anywhere at anytime. While it may give 70% or 80% average accuracy over the whole population, it still may not be suitable for many individuals. There is a clear evidence that prediction and treatment based on such global models are only effective for some of the patients (about 70% at average) (Shabo, 2007) leaving the rest of patients with no effective treatment, and in many cases facing worsening of their condition or even death.

The rationale behind the personalised modelling paradigm is that since each person is different, the most effective treatment could be only achieved if it is based on the detailed analysis of data available for this particular patient. With the advancement of science and technology, it is now possible to obtain and utilise a wide range of personal data such as: DNA, RNA, gene and protein expression, clinical tests, age, gender, BMI, inheritance, food and drug intake, disease, ethnicity, etc. (Shabo, 2007; Hindorff et al., 2009; WTCCC, 2007).

The goal is to create an accurate personalised computational model using information for an individual and the available information for other individuals that is related to the same problem. Achieving a higher accuracy of prediction of a personalised risk for a disease or the effect of treatment may mean saving millions of lives, significantly reducing the cost for treatment, and improving the quality of life of hundreds of millions of patients.

The available methods for personalised modelling do not solve the task completely as they optimise only partially a model for an individual (Nevins et al., 2003; Kasabov et al., 2008b; Song and Kasabov, 2005, 2006). These methods are usually derivatives of the K-Nearest Neighbour method (K-NN), where for a pre-defined set of variables describing an individual with unknown outcome and a population of individuals with known outcomes, the closest K samples to the new one are selected from the population data forming a neighbourhood. The outcome for the new sample is decided based on the majority outcomes in the neighbourhood. Modifications of the K-NN method include WKNN (Vapnik, 1998), WWKNN (Kasabov, 2009, 2007b), TWNFI (Song and Kasabov, 2005, 2006; Kasabov, 2007a).

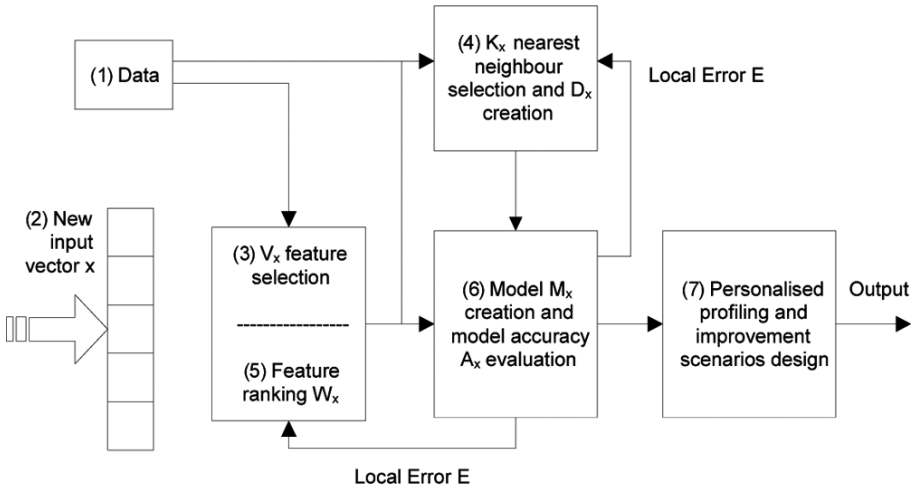
The above methods are suitable only for the problems defined by a small set of variables. In reality, personalised data usually includes thousands of genes, proteins, SNPs, clinical, demographic and other variables. However, using the complete set of available variables would be detrimental to the modelling results, as most of the variables would be redundant. Pre-selecting a set of variables based on their

statistical significance for the whole population space may not be appropriate either, as variables' importance varies depending on the particular sub-space of the problem space (Vapnik, 1998; Kasabov, 2007b). An efficient diagnosis and treatment of a person would require the creation of their personalised profile based on the important variables within the person's sub-space of neighbouring samples. The selection of the neighbourhood of closest samples depends on the selected variables. The overall efficiency of the classification/prediction model would depend on the integrated optimisation of variables, neighbourhood data and parameters of the model, in their concert. Here we present a new IMPM, its implementation and some experimental results for three types of medical decision support problems.

2 An Integrated Method for Personalised Modelling (IMPM) utilising features, model parameters and neighbourhood optimisation

The proposed IMPM method is developed based on the following strategy. For every new individual sample (new input vector) all aspects of their personalised model (variables, neighbouring samples, type of models and model parameters), are optimised together using the accuracy of the outcome achieved for the local neighbourhood of the sample as an optimisation criterion. Next, a personalised model and personalised profile are derived that use the selected variables and the neighbouring samples with known outcomes. The sample's profile is compared with average profiles of the other outcome classes in the neighbourhood (e.g., good outcome, or bad outcome of disease or treatment). The difference between the points and average profiles based on important variables that may need to be modified through treatment. A functional block diagram of the proposed IMPM is shown in Figure 1.

Figure 1 A functional block diagram of the proposed IMPM



Source: Adapted from Kasabov (2008)

2.1 A detailed description of the personalised modelling method

The proposed method consists of the following procedures (Kasabov, 2008):

P1 Data collection, data filtering, storage and update.

P2 Compiling the input vector x for a new patient.

P3 Selecting a subset of relevant to the new sample x variables (features) V_x from a global variable set V.

P4 Selecting a number K_x of samples from the global data set D and forming a neighbourhood D_x of similar samples to x using the variables from V_x to define the similarity.

P5 Ranking the V_x variables within the local neighbourhood D_x in order of importance to the outcome, obtaining a weight vector W_x.

P6 Training and optimising a local prognostic/classification model M_x, that has a set of model parameters P_x, a set of variables V_x and local training/testing data set D_x.

P7 Generating a functional profile F_x for the person x using the selected set V_x of variables, along with the verage profiles of the samples from D_x that belong to different outcome classes, e.g., F_i and F_j. Performing a comparative analysis between F_x, F_i and F_j to define what variables from V_x are the most important for the person x that make him or her very differential from the desired class. These variables may be used to define a personalised course of treatment.

Procedures P3–P6 are repeated a number of iterations or until a desired local accuracy of the model for a local data set D_x is achieved. The optimisation of the parameters of the personalised model V_x, K_x and D_x is global and is achieved through multiple runs of a Genetic Algorithm (GA) that is a type of evolutionary algorithm (Goldberg, 1989; Kasabov, 2007a; Hu and Kasabov, 2009; Mohan and Kasabov, 2005). The resulting competing personalised models for x form a population of such models that are evaluated over iterations (generations) using a fitness criterion – the best accuracy of outcome prognosis for the local neighbourhood of x. Operators of crossover and mutation are applied in the search for the best local model (refer to Figure 2). All variables and parameters of the personalised model form an integrated single 'chromosome' (refer to Figure 3) where variable values are optimised together as a global optimisation.

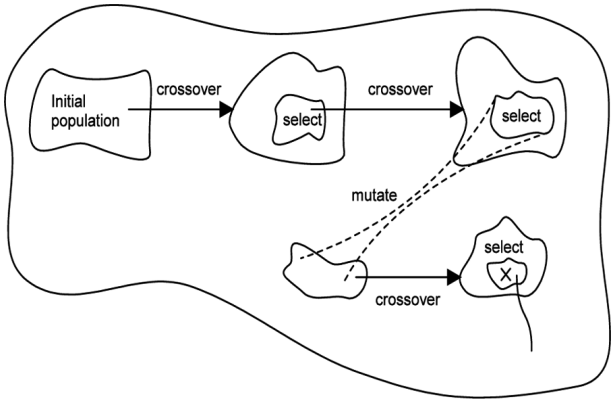
Initially, it is assumed that all variables from a set V have equal absolute and relative importance for a new sample x in relation to predicting its unknown output y:

$$w_{v1} = w_{v2} =, \ldots, = w_{vq} = 1 \tag{1}$$

and

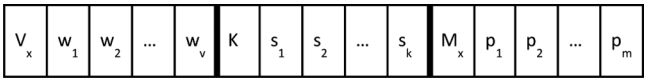
$$w_{v1,\mathrm{norm}} = w_{v2,\mathrm{norm}} =, \ldots, = w_{vq,\mathrm{norm}} = 1/q. \tag{2}$$

Figure 2 A schematic diagram of a GA operation



Source: Adapted from Kasabov (2007a)

Figure 3 A chromosome for an evolutionary computation based integrated, global optimisation of the following parameters ('genes'): number of selected variables V_x; their corresponding weights W_x; number K of nearest neighbours to x; set of selected K samples $s_1 - s_K$ forming a data subset D_x; local prognostic model M_x (e.g., classification algorithm); set of parameters P_m for the M_x (e.g., classification threshold)



The initial numbers for the variables V_x and K_x may be determined in a variety of different ways without departing from the scope of the method. For example V_x and K_x may be initially determined by an assessment of the global data set in terms of size and/or distribution of the data. Minimum and maximum values of these parameters may also be established based on the available data and the problem analysis. For example, $V_{x\_\min} = 3$ (minimum three variables used in a personalised model) and $V_{x\_\max} < K_x$ (the maximum variables used in a personalised model is not larger than the number of samples in the neighbourhood D_x of x), usually $V_{x\_\max} < 20$. The initial set of variables may include expert knowledge, i.e., variables which are referenced in the literature as highly correlated to the outcome of the problem (disease) in a general sense (over the whole population). Such variables for example are the BRCA genes, when the problem is predicting outcome of breast cancer (van't Veer et al., 2002). For an individual patient the BRCA genes may interact with some other genes, which interaction will be specific for the person or a group of people and is likely to be discovered through local or/and personalised modelling only (Kasabov, 2007b).

A major advantage of IMPM is that the modelling process can start with all relevant variables available for a person, rather than with a pre-fixed set of variables in a global model, when compared with global or local modelling. Such a global model may well be statistically representative for a whole population, but not necessarily representative for a single person in terms of optimal model and best profiling and prognosis for this person.

Selecting the initial number K_x of neighbouring samples and the minimum and the maximum numbers $K_{x\_\min}$ and $K_{x\_\max}$ will also depend on the data available and on the problem in hand. A general requirement is that $K_{x\_\min} > V_x$, and, $K_{x\_\max} < cN$, where c is a ratio, e.g., 0.5, and N is the number of samples in the neighbourhood D_x of x. Several formulas have been already suggested and experimented (Vapnik, 1998), e.g.,:

- $K_{x\_\min}$ equals the number of samples that belong to the class with a smaller number of samples when the data is imbalanced (one class has many more samples, e.g., 90%, than the another class) and the available data set D is of small or medium size (e.g., hundreds of samples)
- $K_{x\_\min} = \sqrt{N}$, where N is the total number of samples in the data set D.

At subsequent iterations of the method, the parameters V_x and K_x along with all other parameters are optimised via an optimisation procedure such as:

- exhaustive search, where all or some possible values of all or some of the parameters V_x, W_x, K_x, M_x and P_x are used in their combination and the model M_x with the best accuracy is selected
- an evolutionary algorithm, such as GA (Goldberg, 1989), optimises all or some parameters that form the 'chromosome' from Figure 3.

The closest K_x neighbouring vectors to x from D are selected to form a new data set D_x. A local weighted variable distance measure is used to weigh the importance of each variable V_l $(l = 1, 2, \ldots, q)$ to the accuracy of the model outcome calculation for all data samples in the neighbourhood D_x. For example, the distance between x and z from D_x is measured as a local weighted variable distance:

$$d_{x,z} = \frac{\sqrt{\sum_{l=1}^{q}(1 - w_{l,\mathrm{norm}})(x_l - z_l)^2}}{q} \tag{3}$$

where w_l is the weight assigned to the variable V_l and its normalised value is calculated as:

$$w_{l,\mathrm{norm}} = \frac{w_l}{\sum_{i=1}^{q} w_i}. \tag{4}$$

Here the distance between a cluster centre (in our case it is the vector x) and cluster members (data samples from D_x) is calculated not only based on the geometrical distance, as it is in the traditional nearest neighbour methods, but on the relative variable importance weight vector W_x in the neighbourhood D_x as suggested in Kasabov (2007b). After a subset D_x of K_x data samples are selected based on the variables from V_x, the variables are ranked in a descending order of their importance for prediction of the output y of the input vector x and a weighting vector W_x is obtained. Through an iterative optimisation procedure the number of the variables V_x to be used for an optimised personalised model M_x will be reduced, selecting only the most appropriate variables that will provide the best personalised prediction accuracy of the model M_x. For the weighting W_x (i.e., ranking) of the V_x

variables, alternative algorithms can be used, such as t-test, Signal-to-Noise Ratio (SNR), etc.

In the SNR algorithm, W_x are calculated as normalised coefficients and the variables are sorted in descending order: $V_1, V_2, \ldots, V_v$, where $w_1 >= w_2 >= \ldots >= w_v$, calculated as follows:

$$w_l = \frac{|M_l^{\text{class1}} - M_l^{\text{class2}}|}{std_l^{\text{class1}} + std_l^{\text{class2}}} \tag{5}$$

where M_l^{classs} and std_l^{classs} are respectively the mean value and the standard deviation of variable x_l for all vectors in D_x that belong to class s. This method is very fast, but evaluates the importance of the variables in the neighbourhood D_x one by one and does not take into account a possible interaction between the variables, which might affect the model output.

A classification or prediction learning procedure is applied to the neighbourhood D_x of K_x data samples to derive a personalised model M_x using the already defined variables V_x, variable weights W_x and a model parameter set P_x. A number of different classification or prediction procedures can be used such as: KNN, WKNN, WWKNN (Kasabov, 2007b), MLR, SVM, TWNFI (Song and Kasabov, 2006), and others. In the Weighted KNN (WKNN) classification model, the outcome for the new sample is calculated based on the weighted outcomes of the individuals in the neighbourhood according to their distance to the new sample. In the WWKNN model (Kasabov, 2007b) variables are ranked and weighted according to their importance for separating the samples of different classes in the neighbourhood area in addition to the weighting according to the distance as in WKNN. In the TWNFI model – transductive, weighted neuro-fuzzy inference system (Song and Kasabov, 2006), the number of variables in all personalised models is fixed, but the neighbouring samples used to train the personalised neuro-fuzzy classification model are selected based on the variable weighted distance to the new sample as it is in the WWKNN.

When using the WWKNN method (Kasabov, 2007b), the output value y for the input vector x is calculated using the formula:

$$y = \frac{\sum_{j=1}^{K} a_j y_j}{\sum_{j=1}^{K} w_j} \tag{6}$$

where y_j is the output value for the sample x_j in the neighbourhood D_x of x and:

$$a_j = \max(d) - [d_j - \min(d)] \tag{7}$$

In equation (7), the vector distance $d = [d_1, d_2, \ldots, d_K]$ is defined as the distances between the new input vector x and the nearest samples (x_j, y_j) for $j = 1$ to K_x; $\max(d)$ and $\min(d)$ are the maximum and minimum values in d respectively. Euclidean distance d_j between vector x and a neighbouring one x_j is calculated as:

$$d_j = \sqrt{\sum_{l=1}^{V} (1 - w_l)(x_l - x_{jl})^2} \tag{8}$$

where w_l is the coefficient weighing variable x_l in the neighbourhood D_x of x (e.g., w_l can be calculated by a SNR algorithm, refer to equation (5).

When using the TWNFI classification or prediction model (Song and Kasabov, 2006), the output y for the input vector x is calculated as follows:

$$y = \frac{\sum_{l=1}^{m} \frac{n_l}{\delta_l^2} \prod_{j=1}^{P} \alpha_{lj} \cdot \exp\left[- \frac{w_j^2 (x_{ij} - m_{lj})^2}{2\sigma_{lj}^2} \right]}{\sum_{l=1}^{m} \frac{1}{\delta_l^2} \prod_{j=1}^{P} \alpha_{lj} \cdot \exp\left[- \frac{w_j^2 (x_{ij} - m_{lj})^2}{2\sigma_{lj}^2} \right]} \tag{9}$$

where m is the number of the closest clusters to the new input vector x; each cluster l is defined as a Gaussian function G_l in a V_x dimensional space with a mean value m_l as a vector and a standard deviation δ_l as a vector too; $x = (x_1, x_2, \ldots, x_v)$; α_l (also a vector across all variables V) is membership degree to which the input vector x belongs to the cluster Gaussian function G_l; n_l is a parameter of each cluster (Song and Kasabov, 2006). The detailed algorithm of TWNFI is described in Appendix.

A local accuracy (local error E_x), that estimates the personalised accuracy of the personalised prognosis (classification) for the data set D_x using model M_x is evaluated. This error is a local one, calculated in the neighbourhood D_x, rather than a global accuracy, that is commonly calculated for the whole problem space D. A variety of methods for calculating error can be employed such as:

- RMSE (root-mean square error)
- AUC (area under the receiving operating characteristic curve)
- AE (absolute error).

We propose here another formula for calculating local error that can be used for model optimisation:

$$E_x = \frac{\sum_{j=1}^{K_x} (1 - d_{xj}) \cdot E_j}{K_x} \tag{10}$$

where d_{xj} is the weighted Euclidean distance between sample x and sample S_j from D_x that takes into account the variable weights W_x (see equation (3)); E_j is the error between what the model M_x calculates for the sample S_j from D_x and what its real output value is.

In the above formula the closer a data sample S_j to x is, based on a weighted distance measure, the higher its contribution to the error E_x will be. The calculated personalised model M_x accuracy is:

$$A_x = 1 - E_x. \tag{11}$$

The best accuracy model obtained is stored for a future improvement and optimisation purposes. The optimisation procedure iteratively returns to all previous procedures (P2–P6) to select another set of parameter values for the parameter vector (refer to Figure 3), according to one of the optimisation

procedures listed above (exhaustive search, GA, a combination between them) until the model M_x with the best local accuracy is achieved. The method also optimises parameters P_x of the classification/prediction procedure. Once the best model M_x is derived, an output value y for the new input vector x is calculated using this model. After the output value y for the new input vector x is calculated a personalised profile F_x of the person represented as input vector x is derived, assessed against possible desired outcomes for the scenario, and possible ways to achieve an improved outcome will be designed, which is also a major novelty of this method. A personal improvement scenario consists of suggested changes in the values of the person's variables to improve the outcome for x is designed. The x profile F_x is formed as a vector:

$$F_x = \{V_x, W_x, K_x, D_x, M_x, P_x, t\} \tag{12}$$

where the variable t represents the time of the model M_x creation. At a future time $(t + \Delta t)$ the person's input data will change to x^* (due to changes in variables such as age, weight, protein expression values, etc.), or the data samples in the data set D may be updated and new data samples added. A new profile F_x^* derived at time $(t + \Delta t)$ may be different from the current one F_x.

The average profile F_i for every class C_i in the data D_x is a vector containing the average values of each variable of all samples in D_x from class C_i. The importance of each variable (feature) is indicated by its weighting in the weight vector W_x. The weighted distance from the person's profile F_x to the average class profile F_i (for each class i) is defined as:

$$D(F_x, F_i) = \sum_{l=1}^{v} |V_{lx} - V_{li}| \cdot w_l \tag{13}$$

where w_l is the weight of the variable V_l calculated for dataset D_x (see equation (4)).

Assuming that F_d is the desired profile (e.g., normal outcome), the weighted distance $D(F_x, F_d)$ will be calculated as an aggregated indication of how much a person's profile should change to reach the average desired profile F_d:

$$D(F_x, F_d) = \sum_{l=1}^{v} |V_{lx} - V_{ld}| \cdot w_l. \tag{14}$$

A scenario for a person's improvement through changes made to variables (features) towards the desired average profile F_d can be produced as a vector of required variable changes, defined as:

$$\Delta F_{x,d} = \Delta V_{lx,d} \mid l = 1, \ldots, v \tag{15}$$

$$\Delta V_{lx,d} = |V_{lx} - V_{ld}|, \text{ with an importance of } w_l. \tag{16}$$

In order to find a smaller number of variables, as global markers that can be applied to the whole population X, procedures P2–P7 are repeated for every individual x. All variables from the derived sets V_x are then ranked based on their

likelihood to be selected for all samples. The top m variables (most frequently used for individual models) are selected as a set of global set of markers V_m. The procedures P1–P7 will be applied again with the use of V_m as initial variable set (instead of using the whole initial set V of variables). In this case personalised models and profiles are obtained within a set of variable markers V_m that would make treatment and drug design more universal across the whole population X.

3 Results of using the IMPM on case studies for medical decision support

3.1 Software implementation of the IMPM

The IMPM was implemented as a software toolbox, which employed evolutionary computational techniques. We used a coevolutionary based algorithm to search optimised personalised models (Hu and Kasabov, 2009). We have applied our method on three case studies to illustrate the possible uses of the proposed method.

3.2 Personalised modelling with IMPM for colon cancer diagnosis and profiling on gene expression data

The first case study is the personalised modelling for diagnosis and profiling of cancer. A benchmark colon cancer gene expression dataset is used (Alon et al., 1999). It consists of 62 samples, 40 collected from colon cancer patients and 22 from control subjects. Each sample is represented by 2000 gene expression variables. The objective is to create a diagnostic (classification) system that not only provides an accurate diagnosis, but also profiles the person to help define the best treatment.

An example of a personalised model of colon cancer diagnosis and profiling of a randomly selected person is given in Figure 4–9. Figure 4 shows the evolution (GA) process of feature selection specifically for sample#32 from the colon cancer data through 600 generations. IMPM selects 18 genes (features) out of 2000 genes based the result from the GA optimisation. Figure 5 illustrates the weighted importance of the selected 18 genes in Figure 4. The weighted importance is calculated by a weighted SNR model (refer to equations (4)–(5)). The larger the importance value, the more informative the gene is.

Using the proposed IMPM, an optimised personalised model M_x for sample#32 from the colon cancer data is created. This personalised model M_x consists of the selected 18 informative genes, along with two parameters – classification threshold ($\theta = 0.40$) and the number of neighbouring samples ($K = 18$) are optimised specifically for sample#32. Figure 6 shows the data subset D_x with 18 samples (the neighbourhood with an appropriate size) of sample#32 using top 3 selected genes (gene 377, 1285 and 1892). These neighbouring samples are:

$$61, 41, 12, 1, 38, 22, 26, 31, 34, 28, 19, 44, 6, 49, 57, 3, 8, 43.$$

The predicted outcome computed by the optimised personalised model M_x is 0.51, which successfully classifies sample#32 into diseased class (class 2) (the classification threshold θ is optimised to 0.40 as a model parameter).

Using the IMPM, a profile and a scenario of potential genome improvement for colon sample#32 was created shown in Figure 7. Desired average profile is the

Figure 4 The evolution of feature (variable) selection for sample#32 from the Colon cancer data (600 generations of GA optimisation; the lighter the colour, the higher the probability of the feature to be selected; each feature is represented as one bit on the horizontal axis; at the beginning all features are assigned equal probability to selected as 0.5) (see online version for colours)

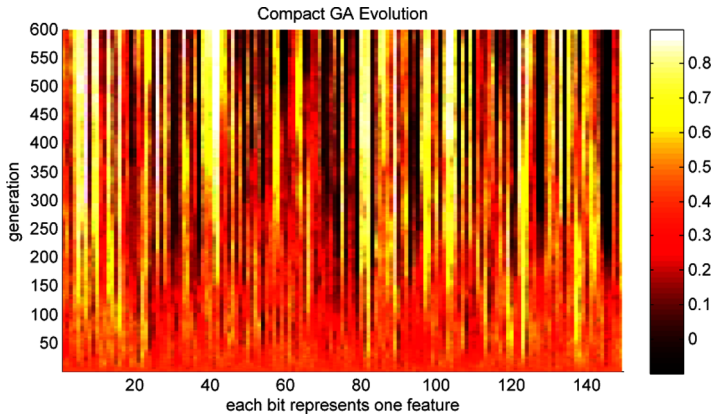
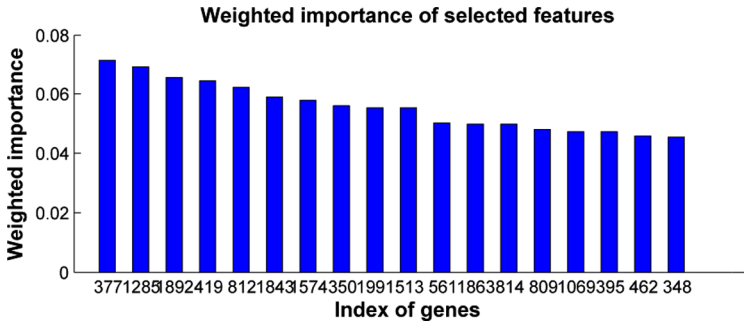


Figure 5 The weighted importance of the selected features for sample#32 using weighted SNR based model (refer to equation (4)–(5)) (see online version for colours)



average gene expression level from healthy samples group and desired improvement value identifies the change of the gene expression level that this patient (sample#32) should follow in order to recover from the disease. For example, the expression level of gene 377 of sample#32 is 761.3790, while the average class profile for class 1 (normal class) and class 2 (diseased class) are: *233.8870* (for class 1) and *432.6468* (for class 2). The distance between the gene expression level of gene 377 for sample#32 and the desired average class 1 profile is 527.4920, i.e., a potential solution can be given to the colon cancer patient (sample#32) to decrease his or her gene 377 expression level from 761.3790 to 233.8870. The information in the generated profile can be used for designing personalised treatment for cancer patients.

To find a small number of variables (potential markers) for the whole population of colon cancer data, we have used the approach as follows: Based on the experiment result for every sample, we selected 20 most frequently used genes as potential global markers. Table 1 lists these 20 global markers with their biological

Figure 6 Sample#32 (the blue dot) is plotted with its 18 neighbouring samples selected by IMPM (red triangles – cancer samples and green triangles – control) in the 3D space of the top 3 gene variables (genes 377, 1285 and 1892) from Figure 5 (see online version for colours)

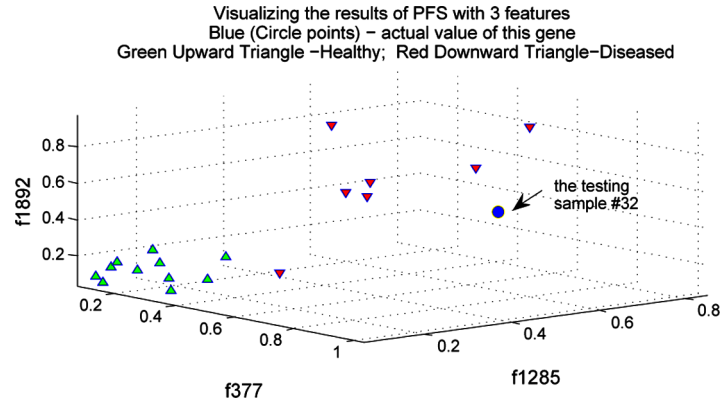


Figure 7 The profile of sample#32 (blue dots) vs. the average local profile of the control (green) and cancer (red) samples using the 18 selected genes from Figure 5 as derived through the IMPM (see online version for colours)

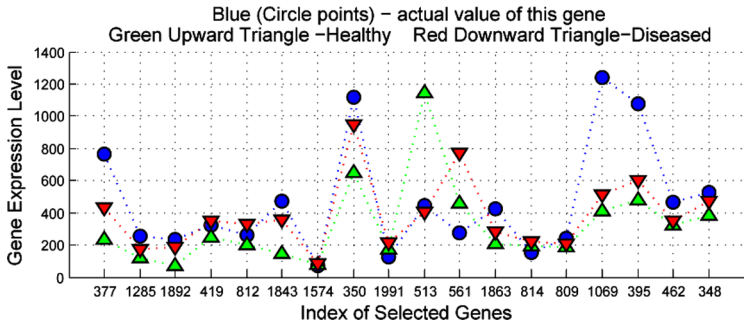
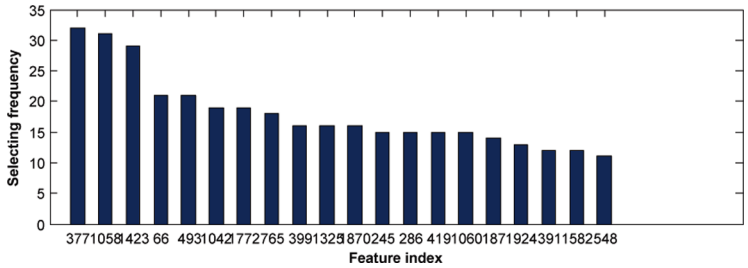
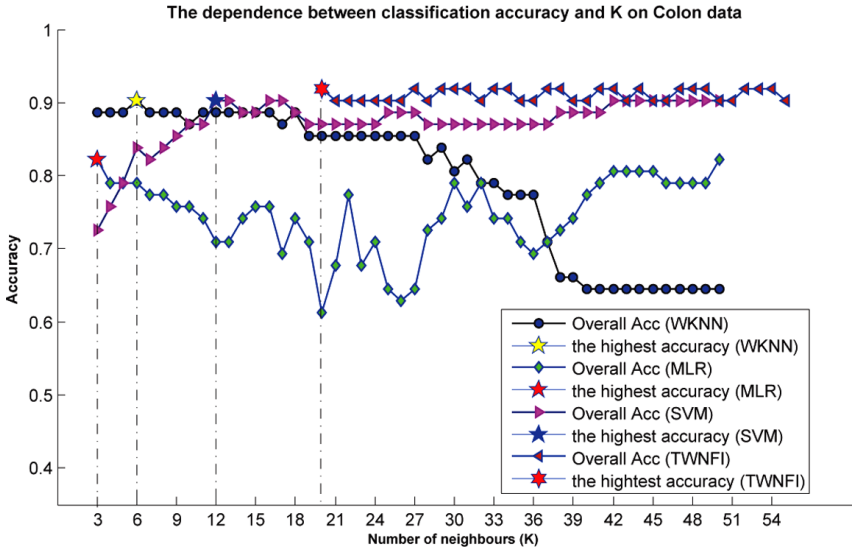


Figure 8 The 20 most frequently selected genes using IMPM across all colon cancer data samples, where x axis represents the index of the gene in the data and y axis is the frequency of the gene as the marker of the optimised personalised models for which this gene has been selected (see online version for colours)



information. Here we use 20 selected genes as global markers. The number of 20 is based on the suggestion in Alon's work (Alon et al., 1999).

Figure 9 A comparison of classification results obtained by 4 classification algorithms using 20 potential maker genes from Figure 8 , where x axis represents the size of neighbourhood and y axis is the average classification accuracy across all samples. The best accuracy is obtained with the use of the TWNFI classification algorithm (91.90%) (see online version for colours)



The next objective of our experiment is to investigate whether utilising these 20 potential marker genes can lead to improved colon cancer classification accuracy and what classification algorithm will perform best in the proposed IMPM. Four classification algorithms are tested as personalised models in this experiment, including WKNN, MLR, SVM and TWNFI. All the classification results from four classifiers are validated based on leave-one-out cross validation (LOOCV) across the whole dataset. Figure 9 shows the average accuracy obtained by these four algorithms with different size (K_x) of neighbourhood. Table 2 summarises the classification results from the four classification algorithms using 20 selected potential marker genes. WKNN and a localised SVM yielded improved classification accuracy (90.3%) when compared to the global model (Alon et al., 1999). However, the TWNFI classifier obtained the best classification performance (91.9%). Our results suggest that a small set of marker genes selected by the IMPM could lead to improved cancer classification accuracy.

3.3 Crohn's disease risk evaluation on geneome SNPs data

The second case study is personalised modelling for risk of Crohn's Disease (CD) risk evaluation based on whole genome SNPs data. Large repositories of SNPs data have been collected from control patients and ill patients (Hindorff et al., 2009; WTCCC, 2007). The data can be utilised to accurately predict an individual's risk of disease based on a personalised DNA profiling using the IMPM. The UK's Welcome Trust Case Control Consortium (WTCCC) data was collected as part of a genome-wide association study project of 14,000 cases of 7 major diseases and a shared set of 3000 controls (WTCCC, 2007).

Table 1 The 20 most frequently selected genes (potential marker genes) using the proposed IMPM across all colon cancer gene data samples (see Figure 8)

Index of gene	*GenBank accession number*	*Description of the gene (from GenBank)*
G377	Z50753	H.sapiens mRNA for GCAP-II/ uroguanylin precursor
G1058	M80815	H.sapiens a-L-fucosidase gene, exon 7 and 8, and complete cds
G1423	J02854	Myosin regulatory light chain 2, smooth muscle ISOFORM (HUMAN)
G66	T71025	Human (HUMAN)
G493	R87126	Myosin heavy chain, nonuscle (Gallus gallus)
G1042	R36977	P03001 Transcription factor IIIA
G1772	H08393	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)
G765	M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6
G399	U30825	Human splicing factor SRp30c mRNA, complete cds
G1325	T47377	S-100P PROTEIN (HUMAN)
G1870	H55916	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR (HUMAN)
G245	M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6
G286	H64489	Leukocyte Antigen CD37 (Homo sapiens)
G419	R44418	Nuclear protein (Epstein-barr virus)
G1060	U09564	Human serine kinase mRNA, complete cds
G187	T51023	Heat shock protein HSP 90-BETA (HUMAN)
G1924	H64807	Placental folate transporter (Homo sapiens)
G391	D31885	Human mRNA (KIAA0069) for ORF (novel proetin), partial cds
G1582	X63629	H.sapiens mRNA for p cadherin
G548	T40645	Human Wiskott-Aldrich syndrome (WAS) mRNA, complete cds

Here, we randomly selected a sample (#392) and evaluated it through 20 runs. The proposed IMPM method produced an applausable prediction accuracy: the prediction for sample#392 was always correct through all 20 runs. The average local accuracy for this sample through 20 runs was 82.45%, which was significantly better than the average accuracy from a published global statistical method (e.g., SVM) that was only around 64%. Our experimental results has shown that the IMPM has worked effectively on sample#392, as the computed local accuracy through 20 runs is very stable – ranges from 81% to 83%.

Figure 10 shows the number of selected features for sample#392 in each of the 20 runs using the proposed IMPM method. The selecting frequency of each feature

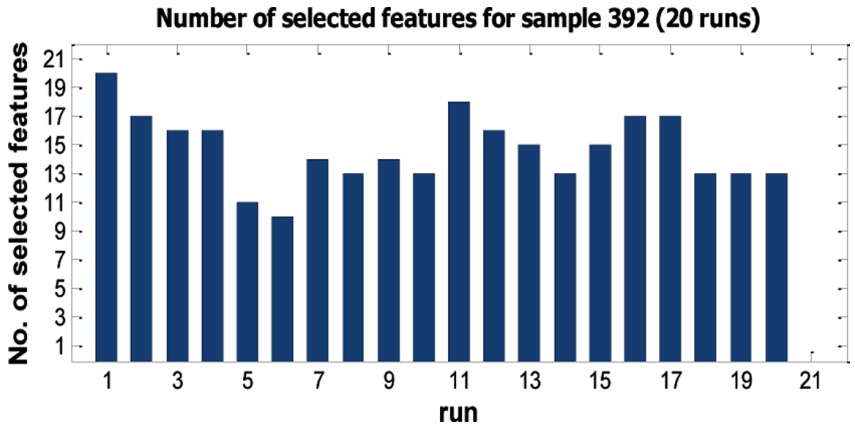
for testing sample#392 through 20 runs is shown in Figure 11. Here *Age* is the most important feature for CD prediction, along with other top 5 selected SNPs:

Feature Id	*SNP Id*	*Selecting frequency* (/20*times*)
20	X4252400_T	19
24	X2155777_T	18
12	X7683921_A	14
9	X2270308_T	13
23	X10883359_G	13

Table 2 The best classification accuracy obtained by four classification algorithms on colon cancer data with 20 potential maker genes. Overall – overall accuracy; Class 1 – class 1 accuracy; Class 2 – class 2 accuracy

Classifier	*Overal* [%]	*Class 1* [%]	*Class 2* [%]	*Neighbourhood size*
MLR (Personalised)	82.3	90.0	68.2	3
SVM (Personalised)	90.3	95.0	81.8	12
WKNN	90.3	95.0	81.8	6
TWNFI	*91.9*	95.0	85.4	20
Original publication (Alon et al., 1999)	87.1	–	–	–

Figure 10 An example of applying the IMPM for personalised modelling risk of Crohns' disease evaluation based on the UK WTCCC data (WTCCC, 2007). A single sample#392 is randomly selected from the data set. The number of selected features for sample#392 in each of the 20 runs of the method is shown (see online version for colours)



These selected important features, including 2 clinical factors (age and gender) and 5 SNPs may be the potential global markers for CD risk evaluation across whole CD patient population. Such information can be utilised for the personalised treatment and drug design for this specific CD disease research.

Figure 11 An example of personalised modelling for risk of Crohns' disease evaluation based on the UK WTCCC data (WTCCC, 2007). The most frequently selected features for sample#392 after 20 runs (see Figure 10): #1 – Age; #20 – SNP X4252400 T; #24 – SNP X2155777 T; #12 – Gender; #12 – SNP X7683921; #9 – SNP X2270308 and #23 – SNP X10883359 (see online version for colours)

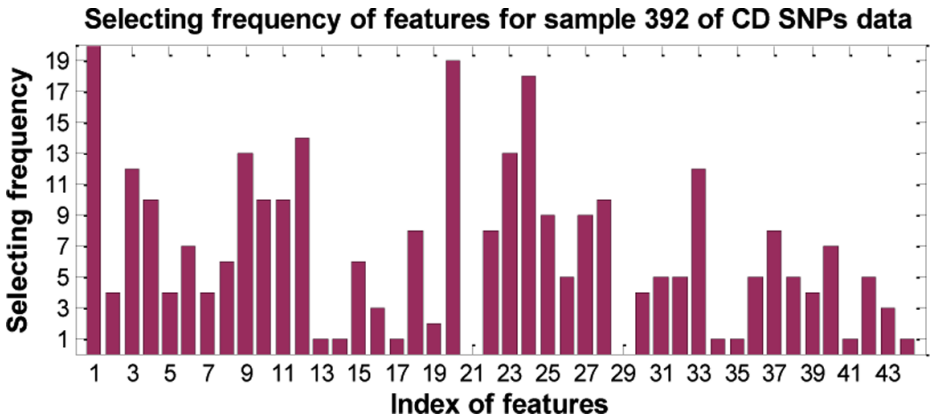
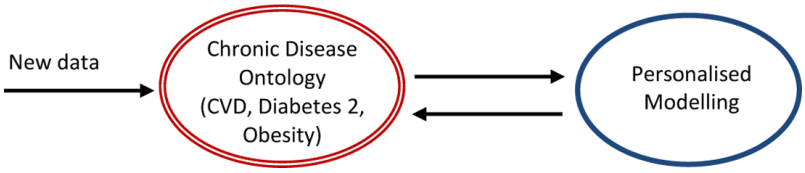


Figure 12 A schematic representation of a personalised modelling system linked to a chronic disease ontology (see online version for colours)



Source: Adapted from Kasabov et al. (2008b)

3.4 Individual risk evaluation of chronic disease

The case study given here only as a general framework, is illustrated in Figure 12. It shows a block diagram of a personalised modelling system for evaluation of individual risks of chronic disease (cardio-vascular disease, diabetes type 2 and obesity) that is linked to a chronic Disease Ontology. The ontology contains DNA-, genetic-, proteomic-, clinical-, nutritional and other types of data of both control patients and patients of a chronic disease (Kasabov et al., 2008b).

With the advancement of personalised data collection techniques, personalised modelling is expected to play a significant role for the understanding of specific personal conditions and for the design of more efficient personalised treatment for patients with neurodegenerative diseases, such as Alzheimer disease, clinical depression, bipolar disease, Parkinson's disease and others (National Center for Biotechnology Information, 2010; Reggia et al., 1999; Benuskova et al., 2006). The IMPM can be linked to existing brain-gene ontology systems as shown in Figure 12. Such ontology is the BGO proposed in the work (Kasabov et al., 2008a; Benuskova et al., 2006).

4 Conclusion

When compared to global or local modelling, the proposed personalised modelling method (IMPM) has a major advantage. In our method, the modelling process starts with all relevant variables available for a person, rather than with a fixed set of variables required by a global model that may well be statistically representative for a whole population, but not necessarily representative for a single person in terms of best prognosis for this person. The proposed IMPM leads to a better prognostic accuracy and a computed personalised profile. With global optimisation, a small set of variables (potential markers) can be identified from the selected variable set across the whole population. This information can be utilised for the development of new more efficient drugs. A scenario for outcome improvement is also created by the IMPM, which can be utilised for the decision of efficient personalised treatment. We hope that this paper will motivate the biomedical applications of personalised modelling research.

Personalised modelling methods and systems are not going to substitute experts and current global or local modelling methods, but they are expected to derive information that is specifically relevant to a person and help individuals and clinicians make better decisions, thus saving lives, improving quality of life, and reducing cost of treatment.

Acknowledgements

This work was supported by the Tertiary Education Commission of New Zealand through Top Achiever Doctoral Scholarship to Yingjie Hu and funded by Auckland University of Technology to Knowledge Engineering and Discovery Research Institute (KEDRI). We acknowledge the technical and editorial assistance from Joyce D'Mello and Diana Kassabova and the help with the SNPs data by Dr. Rod Lea. The presented IMPM is available as a patent description (Kasabov, 2008). The software implemented of the IMPM was developed by Dr. Yingjie Hu, who also conducted the experiments on the case study data. The software is available upon request.

Funding: This work was supported by the Tertiary Education Commission of New Zealand through Top Achiever Doctoral Scholarship to Yingjie Hu.

References

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) 'Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 96, pp.6745–6750.

Benuskova, L., Jain, V., Wysoski, S.G. and Kasabov, N. (2006) 'Computational neurogenetic modeling: a pathway to new discoveries in genetic neuroscience', *Intl. Journal of Neural Systems*, Vol. 16, No. 3, pp.215–227.

Goldberg, D. (1989) *GeneticAlgorithm in Search, Optimization and Machine Learning*, Kluwer Academic, MA.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) 'Potential etiologic and functional implications of genome-wide association loci for human diseases and traits', *Proceedings of the National Academy of Sciences*, Vol. 106, No. 23, pp.9362–9367.

Hu, Y. and Kasabov, N. (2009) 'Coevolutionary method for gene selection and parameter optimization in microarray data analysis', in Leung, C., Lee, M. and Chan, J. (Eds.): *Neural Information Processing*, Springer-Verlag, Berlin/Heidelberg, pp.483–492.

Kasabov, N., Jain, V. and Benuskova, L. (2008a) 'Integrating evolving brain-gene ontology and connectionist-based system for modeling and knowledge discovery', *Neural Networks*, Vol. 21, Nos. 2–3, pp.266–275.

Kasabov, N., Song, Q., Benuskova, L., Gottgtroy, P.C.M., Jain, V., Verma, A., Havukkala, I., Rush, E., Pears, R., Tjahjana, A., Hu, Y. and MacDonell, S.G. (2008b) 'Integrating local and personalised modelling with global ontology knowledge bases for biomedical and bioinformatics decision support', in Smolinski, T.G., Milanova, M.G. and Hassanien, A.E. (Eds.): *Computational Intelligence in Biomedicine and Bioinformatics*, Springer, Berlin, pp.93–116.

Kasabov, N. (2007a) *Evolving Connectionist Systems: The Knowledge Engineering Approach*, Springer, London.

Kasabov, N. (2007b) 'Global, local and personalized modelling and pattern discovery in bioinformatics: an integrated approach', *Pattern Recognition Letters*, Vol. 28, No. 6, pp.673–685.

Kasabov, N. (2008) *Data Analysis and Predictive Systems and Related Methodologies – Personalised Trait Modelling System*, New Zealand Patent No. 572036, PCT/NZ2009/000222, NZ2009/000222-W16-79.

Kasabov, N. (2009) 'Soft computing methods for global, local and personalised modeling and applications in bioinformatics', in Balas, V.E., Fodor, J. and Varkonyi-Koczy, A. (Eds.): *Soft Computing Based Modeling in Intelligent Systems*, Springer, Berlin, Heidelberg, pp.1–17.

Mohan, N. and Kasabov, N. (2005) 'Transductive modeling with ga parameter optimization', *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, Montreal, Vol. 2, July, pp.839—844.

National Center for Biotechnology Information (2010) National Center for Biotechnology Information (US), Gene and Disease, November 2010, http://www.ncbi.nlm.nih.gov/books/NBK22183/

Nevins, J.R., Huang, E.S., Dressman, H., Pittman, J., Huang, A.T. and West, M. (2003) 'Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction', *Human Molecular Genetics*, Vol. 12, No. 2, pp.R153–R157.

Reggia, J.A., Ruppin, E. and Glanzman, D. (1999) *Disorders of Brain, Behaviour and Cognition: The Neurocomputational Perspective*. Elsevier, New York.

Shabo, A. (2007) 'Health record banks: integrating clinical and genomic data into patient-centric longitudinal and cross-institutional health records', *Personalised Medicine*, Vol. 4, No. 4, pp.453–455.

Song, Q. and Kasabov, N. (2005) 'Nfi: a neuro-fuzzy inference method for transductive reasoning', *Fuzzy Systems, IEEE Transactions on*, Vol. 13, No. 6, pp.799–808.

Song, Q. and Kasabov, N. (2006) 'Twnfi – a transductive neuro-fuzzy inference system with weighted data normalization for personalized modeling', *Neural Networks*, Vol. 19, No. 10, pp.1591—1596.

van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002) 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature*, Vol. 415, No. 6871, pp.530–536.

Vapnik, V.N. (1998) *Statistical Learning Theory*, Wiley, New York.

WTCCC (2007) 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature*, Vol. 447, No. 7145, pp.661–678.

Appendix: TWNFI – a transductive neuro-fuzzy inference system with weighted data normalisation for personalised modelling

TWNFI (Song and Kasabov, 2006) is a dynamic neuro-fuzzy inference system in which a local model is created for analysing each new data vector x_v. A basic block diagram of TWNFI is illustrated in Figure 13.

Giving a training dataset X, for each new data vector x_v, TWNFI creates a unique model with the application of the following steps (Song and Kasabov, 2006):

1 *Normalisation*:

- normalise the training data X and the new data vector x_v (values range from 0 to 1)
- initialise the weights of every input variables (features) to 1.

2 Identifying an appropriate neighbourhood (D_v) for x_v;
Find N_v samples from training data that are closest to x_v based on the weighted normalised Euclidean distance calculated as:

$$\|x - y\| = \sqrt{\frac{\sum_{j=1}^{P} w_j (x_j - y_j)^2}{P}} \tag{17}$$

where x_j and y_j are two vectors in the given problem space, N is the number of samples, and w is a weight vector.

3 Calculate the distance $d_i, i = 1, \ldots, N_v$ using equation (17). d_i is the distance between each sample in D_q and x_v. Each sample's weight is calculated as:

$$v_i = 1 - (d_i - min(d)), i = 1, 2, \ldots, N_v,$$

where $\min(d)$ is the minimum number of elements in the distance vector $d = [d_1, d_2, \ldots, d_{N_v}]$.

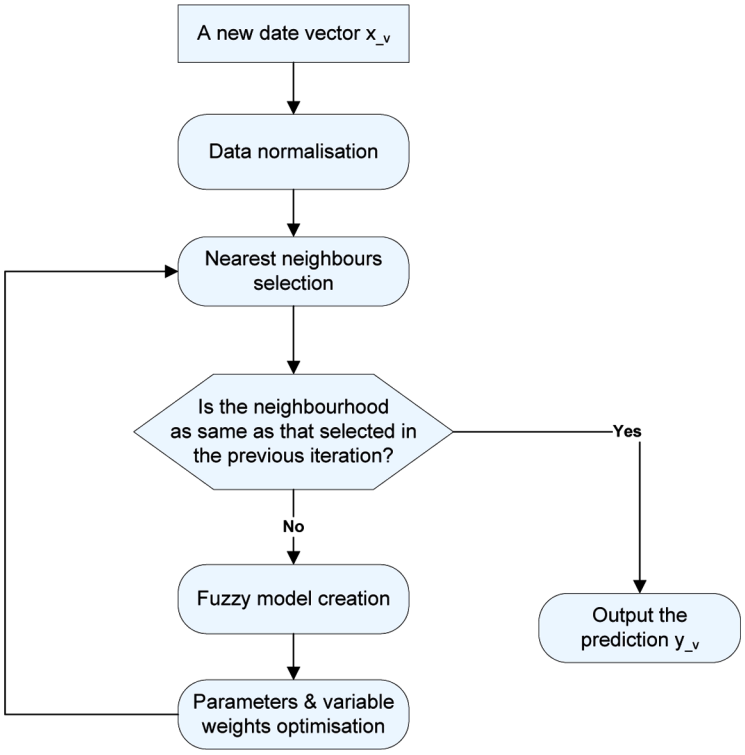
4 Cluster and partition the input subspace that consists of N_v selected training samples; Create fuzzy rules and set their initial parameter values based on the clustering results. Every fuzzy rule is created as:
the centroid of a cluster is the center of the fuzzy membership function (e.g., a Gaussian membership function) and the cluster radius is taken as the width.

5 Apply the steepest descent approach (back-propagation) to optimise the weights and the parameters of the fuzzy rules in a local model M_v.

6 Find a new set of N_v samples (D_v^*) nearest to x_v (*Step 2*):
if the same samples are found as in the last search, the algorithm goes to the next step;
otherwise, it repeats from *Step 3*.

7 Output the prediction y_v for the new data vector x_v using fuzzy inference on the set of fuzzy rules that constitute the local model M_v.

Figure 13 A basic block diagram of TWNFI (see online version for colours)



Source: Adapted from Song and Kasabov (2006)

The weight and parameters can be optimised as follows: Consider a system having P inputs, one output and M fuzzy rules initially defined by a clustering algorithm, and the lth rule is formed as:
R_l: if x_l is F_{l1} and x_2 is F_{l2} and $\cdots$ x_p is F_{lp}, then y is G_l,
where F_{lj} are the fuzzy sets defined by the following Gaussian membership function:

$$\text{Gaussian } MF = \alpha \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \tag{18}$$

and G_l can be defined as:

$$\text{Gaussian } MF = \exp\left(-\frac{(y-n)^2}{2\delta^2}\right). \tag{19}$$

Thus, the output of the system for an input vector $x_i = [x_1, x_2, \ldots, x_p]$ can be calculated by a modified centre average defuzzification function as:

$$f(x_i) = \frac{\sum_{l=1}^{M} \frac{n_l}{\delta_l^2} \prod_{j=1}^{P} \alpha_{lj} \exp[-\frac{w_j^2 (x_{ij} - mlj)^2}{2\sigma_{lj}^2}]}{\sum_{l=1}^{M} \frac{1}{\delta_l^2} \prod_{j=1}^{P} \alpha_{lj} \exp[-\frac{w_j^2 (x_{ij} - mlj)^2}{2\sigma_{lj}^2}]} \tag{20}$$

where, w_j is the current weight vector for the input variables and n_l is the point having maximum membership value in the $l\mathrm{th}$ output set).