Neural Networks Based System for Cancer Diagnosis Support

Witold Jacak and Karin Pröll

Dept. of Software Engineering at Hagenberg
Upper Austrian University of Applied Sciences
Softwarepark 11, A 4232 Hagenberg, Austria
jacak@fh-hagenberg.at proell@fh-hagenberg.at

Abstract. The paper presents the analysis of two different approaches for a system to support cancer diagnosis. The first one uses only tumor marker data containig missing values to predict cancer occurrence and the second one also includes standard blood parameters. Both systems are based on several heterogeneous artificial neural networks for estimating missing values of tumor markers and they finally caluculate possibilities of different tumor diseases.
Keywords: neural network, tumor marker prediction, cancer diagnosis support

1 Introduction

Tumor markers are substances produced by cells of the body in response to cancerous but also to noncancerous conditions. They can be found in body liquids like blood or in tissues and can be used for detection, diagnosis and treatment of some types of cancer. For different types of cancer different tumor markers can show abnormal values and the levels of the same tumor marker can be altered in more than one type of cancer. Neural networks and evoutionary algorithms are proven tools for prediction tasks on medical data [3–5]. In this work we present a neural network based system which can be used as support in cancer diagnosis based on tumor marker values and blood parameters from blood examination. A main focus in this work is laid on the problem of missing values in biomedical data as they make training of neural networks difficult. The cancer prediction system is based on data coming from vectors $\boldsymbol{C} = (C_1, \ldots, C_n)$ containg tumor marker values which are frequently incomplete, containg lots of missing values influencing the plausibility of diagnosis prediction. The question raises if it is possible to increase the quality of cancer prediction by using information beyond tumor marker data. The general goal of a data driven cancer prediction system can be expressed as follows:
Construct a data driven cancer diagnosis support system which:

- *maximizes the probability of correct cancer diagnosis (positive and negative)*
- *minimizes the probability of incorrect diagnosis if a cancerous disease exists*

One efficient method for such a system is the synthesis of complex neural networks for prediction of cancer based on tumor markers values. We need many thousand datasets for training and evaluating the neural networks. As mentioned before these datasets contain lots of mssing values. To overcome this problem we additionally make use of datasets containg a whole blood parameter vector $\boldsymbol{P} = (P_1, \ldots, P_m)$ of each patient. Frequently those vectors are incomplete too. For these reasons we link two independently trained neural network systems into one: The first subsystem is trained only with complete or incomplete tumor marker datasets $\boldsymbol{C}$. The second one includes also blood parameters vectors $\boldsymbol{P}$ to support prediction of cancer possibility.

2 Tumor marker values based cancer diagnosis support system

Cancer diagnosis support uses parallely working networks ($Cancer_k$), with the same structure, trained separately for different types of cancer. The input of each ($Cancer_k$) system is the complete or incomplete vector $\boldsymbol{C}$ of tumor marker specific for a chosen type of cancer, and the output represents the possibility (values between 0 and 1) of a cancer disease. Output values of the network system greater than 0,5 are treated as cancer occurrence.
Each ($Cancer_k$) neural networks system consists of four different groups of neural networks (see Figure 1-Layer 1).

- Group of neural networks (C_{net}) for individual marker $C_i, i = 1, .., n$.
- Feed forward neural network ($C_{Group}FF_{net}$) for vector of marker $\boldsymbol{C}$, with complete or incomplete values.
- Pattern recognition neural network ($C_{Group}PR_{net}$) for vector of marker $\boldsymbol{C}$, with complete or incomplete values.
- Cascade-coupled aggregation method for final calculation of cancer plausibility.

2.1 Group of separate neural networks for individual marker ($\boldsymbol{C_{net}}$)

The first group of neural networks contains parallel coupled neural networks, which are individually trained for different tumor markers. Each neural network is a feed forward network with one hidden layer, one input (normalized tumor marker values) and one output (diagnosis: 0 - cancer not occurs (healthy) and 1 - cancer occurs (sick)). The hidden layer has 6 -10 neurons with tan/sigmoid activation functions. The values of markers can belong to four intervals (Classes). The first interval includes all values less than a *Normal Value* of marker, the second interval includes all values between *Normal Value* and *Extreme Normal Value* of markers, the third interval includes values between *Extreme Normal Value* and *Plausible Value* of marker and the fourth interval includes all values greater than *Plausible Value*. The input values of each network for training and testing are normalized using the respective upper bound of *Plausible Value*. Each

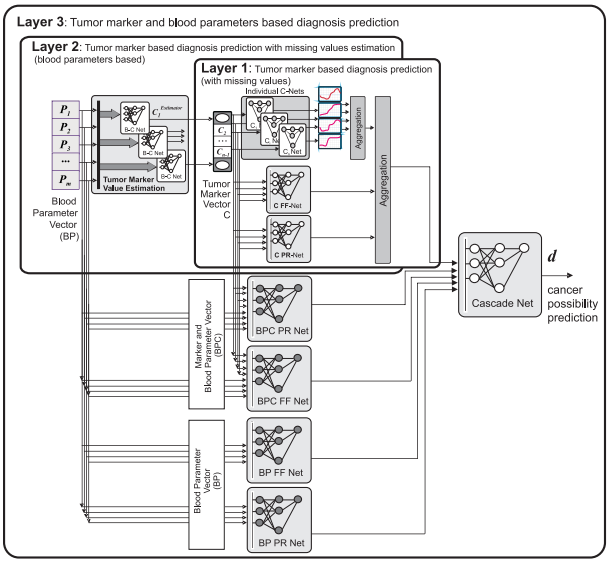

Fig. 1. Structure of $Cancer_k$ prediction system based on tumor marker values

marker value, which is greater than its upper bound, obtains the normalized value 1. For example the threshold point of network trained for all cancer types is 0,46 (75,9 U/ml) and the threshold point of the network trained for breast cancer is 0,52 (85,5 U/ml) for C125 marker. The input of parallel-coupled C_{net} is the vector of tumor marker $\boldsymbol{C} = (C_1, \ldots, C_n)$, where not all C_i values exists. When the tumor marker value in vector $\boldsymbol{C}$ is available, then the adequate C_{net} calculates the predicted cancer possibility. When the marker value in vector $\boldsymbol{C}$ is not available, then the output of C_{net} is set to -1. The individually calculated output values of C_{net} can be aggregated in many different ways. We compare three methods of aggregation:

- Maximum value of all individual network outputs.
  $C_{net}(C) = max\{C^i_{net}(C_i)|i = 1, .., m\}$
- Average value of all individual network outputs, without missing values.
  $C_{net}(C) = avg\{C^i_{net}(C_i)|i = 1, .., m \ \& \ C_i \neq -1\}$
- $Net_{aggregation}$ - neural network trained with individual network outputs (this neural network used fro aggregation can be trained with data containing only one chosen cancer type $Cancer^k$):
  $C_{net}(C) = net_{aggregation}(C^i_{net}(C_i)|i = 1, ..., m)$

We use one aggregation type in the full system. In case of max aggregation: If only one marker of the marker group shows a greater value than the aggregation has yielded this value is taken. The diagnosis prediction based on aggregation of separate cancer predictions of individual marker networks C_{net} is not sufficient for the generalization of cancer occurrence. It is necessary to reinforce the

information coming from data of the whole group of markers. Therefore two neural networks with cumulative marker groups are added. These networks will be trained only for a specific cancer type.

2.2 Feed forward and pattern recognition neural networks for tumor marker group

The vectors $\boldsymbol{C}$ of marker values can again be incomplete. If a tumor marker value in vector $\boldsymbol{C}$ is missing, then this value is set to -1. This allows to generate training sets for a specified cancer type $Cancer_k$ and to train the two neural networks in Figure 1-(Layer 1):

- Feed forward neural network with 16-20 hidden neurons and tansig/linear activation functions ($C_{group}FF_{net}$)
- Pattern recognition network with 16-20 hidden neurons ($C_{group}PR_{net}$).

The outputs of all parallel working networks are coupled into a new vector and this represents an input for the cascade-net ($Cascade_{net}^{k-Cancer}$) for diagnosis generalization with 16 hidden neurons or other aggregation function such as $mean$ or max.

3 Case study: Tumor markers based breast cancer diagnosis support system

3.1 Setup

Training and test datasets were prepared for breast cancer. For the tumor marker group we have taken the C125, C153, C199 and CEA markers ($\boldsymbol{C} = (C_{125}, C_{153}, C_{199}, C_{EA})$). The training and test datasets include about 5100 and 2480 data, respectively. The outputs of individual networks are aggregated with the maximum function and with separately trained perceptron network for breast cancer type. $C_{group}FF_{net}$ and $C_{group}PR_{net}$ networks are trained only for breast cancer. The confusion matrices values between test target data and outputs on test input data for all previously mentioned networks are presented in Table 1. P(1/1) represents the probability estimation of true positives (positive diagnosis and actually cancer disease existent), P(0/1) probability estimation of false negatives (negative diagnosis but actually cancer disease existent), P(1/0) represents the probability estimation of false positives (positive diagnosis and actually cancer disease not existent), P(0/0) represents the probability estimation of true negatives (negative diagnosis and actually cancer disease non existent), $P_{correct}$ represents the probability estimation of all correct diagnoses. All values are percentages. This notation will be used in all of the following tables. The outputs of all parallel working networks are coupled into a new vector and this represents an input for the diagnosis generalization system. This system can be constructed as new cascade-net ($Cascade_{net}^{k-Cancer}$) pattern recognition type, with 16 hidden neurons or as classic aggregation function calculating $mean$ or max values of

coupled first level networks outputs. The confusion matrix between test target data and outputs on test input data for these aggregation methods are presented in Table 1.

Table 1. Confusion Matrix between target diagnosis and predicted diagnosis from different neural networks

Neural Networks	P(1/1)	P(1/0)	P(0/0)	P(0/1)	$P_{correct}$
Individual trained networks for C_i with *max* as aggregation function	8,4	5,2	62,1	24,3	70,5
Individual trained networks for C_i with *perceptron network* as aggregation function	32	40,4	26,9	0,8	58,8
Feed Forward neural network with vector $\boldsymbol{C}$ as input	19,5	12,6	54,7	13,2	74,3
Pattern Recognition neural network with vector $\boldsymbol{C}$ as input	20,5	14,7	56,6	12,2	73,1

All neural networks using aggregation predict cancer with higher quality than individual networks. The coupled system is more pessimistic, it means that the probability of a positive prediction in case no cancerous disease is existent (false positives) is greater than the prediction done with individual networks.

Table 2. Confusion Matrix between target diagnosis and aggregated outputs of parallel working networks

Neural Networks	P(1/1)	P(1/0)	P(0/0)	P(0/1)	$P_{correct}$
Generalized diagnosis prediction with *max* as general aggregation function	23,3	18,4	48,9	9,4	72,2
Generalized diagnosis prediction with *mean* as general aggregation function	20,9	14,9	52,4	11,8	73,3
Generalized diagnosis prediction with *Pattern Recognition* as general aggregation function	16,4	12,5	58	16,3	74,4

4 Tumor markers and blood parameters based cancer diagnosis support system

Missing values in marker data make datasets incomplete, which leads to problems in the training process of neural networks and in consequence to a decrease of quality of diagnosis prediction. Incomplete tumor marker vlaues can be compensated by information coming from values of standard blood parameters (obtained by standard blood examinations) in combination with tumor marker measurement. The vector of blood parameters used as input for the neural networks

can support the training process. The structure of such a system is presented in Figure 1-Layer 2.
The additional information coming from blood parameters examination $\boldsymbol{P}$ can be used to:

- estimation of missing value of tumor markers in vector $\boldsymbol{C}$ and,
- train additional networks for cancer occurrence prediction, which are integrated into one system containg previously described subsystems

Typically 27 blood parameters such as HB, WBC, HKT, MCV, RBC, PLT, KREA, BUN, GT37, ALT, AST, TBIL, CRP, LD37, HS, CNEA, CMOA, CLYA, CEOA, CBAA, CHOL, HDL, CH37, FER, FE, BSG1, TF can be measured during a blood examination in the lab, although the number of parameters examined strongly depends on clinicians needs. For each parameter experimentally defined upper and lower bounds of values are set. We divide the ranges of marker $\boldsymbol{C}$ and blood parametesr P into k non-overlapping classes (as presented before) for normalizing blood parameter and marker values. The system consists of three heterogeneous coupled neural networks working in parallel and a rules based decision-making system for aggregation [1, 2]. The input and output values used in every network for training and testing are normalized using the respective upper bound of Plausible Value. Each value of parameter or marker, which is greater than its upper bound, obtains the normalized value 1. The general system for marker-value estimation contains three neural networks.

- Feed forward neural network (FF) with $\boldsymbol{P}$ inputs (normalized values of blood parameter vectors $\boldsymbol{P}$) and one output, normalized values of marker C_i
- Pattern recognition neural network (PR) with $\boldsymbol{P}$ inputs (normalized values of blood parameter vectors $\boldsymbol{P}$) and k outputs, k-dimensional binary vector coding classes of marker C_i
- Combined feed forward neural network (FC) with $\boldsymbol{P}$ inputs (normalized values of blood parameter vectors $\boldsymbol{P}$) and two outputs: normalized values of marker C_i (as in network FF), and normalized classes of marker C_i.

In this method of value estimation, the quality of estimation will highly depend on missing values in vector $\boldsymbol{C}$. The results of cancer prediction with additionally estimation of missing values of tumor markers in comparison to prediction with incomplete data shows that the probability of incorrect diagnosis of positive cancer appearance decreases (false negatives). Test results (test data include 2400 samples with max. 3 missing values of C-markers) for our case study in breast cancer with the previously shown marker group $\boldsymbol{C}$ for both systems is presented in Table 3. As can be seen the overalll probability of correct diagnoses decreases but also the percentage of false negatives.
All neural networks have one hidden layer and tan/sigmoid transfer function. Empirical testa show that best performance can be obtained using networks with 40 neurons in hidden layer. Neural networks based estimation of marker value lead to introduce four hypotheses x_1, x_2, x_3, x_4 related to the class of tumor marker [2]. For each hypothesis x_1, x_2, x_3, x_4 a plausibility value is calculated too.

Table 3. Confusion matrix of breast tumor prediction based on C124, C153, C199 and CEA marker group without and with blood parameter based marker value estimation

Neural Networks	P(1/1)	P(1/0)	P(0/0)	P(0/1)	$P_{correct}$
Neural networks system without missing value estimation (vector $\boldsymbol{C}$ as input)	16,4	9,3	58	16,3	74,4
Neural networks system with blood parameters based missing value estimation (vector $\boldsymbol{C}$ as input)	21,8	18,9	48,4	10,9	70,2

These hypotheses should be verified to find the maximal probability of tumor value prediction [1]. It can be expected that not all markers can be predicted with high quality. The examples of regression between blood parameters test data and marker value estimation for two tumor markers C153 (regression 0,71) and CEA (regression 0,53) are presented in Figure 2.

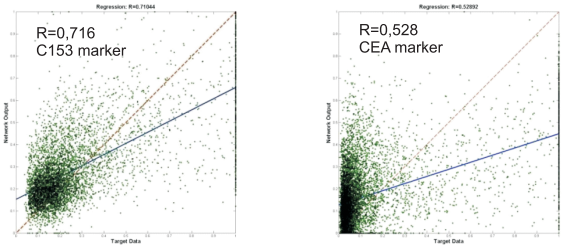

Fig. 2. Regression between test data and predicted tumor marker values for markers C153 (left) and CEA (right)

4.1 System with additional networks for cancer occurrence prediction trained with blood parameter data

As we have observed the introduction of additional information (estimation of missing values) to the system leads to a better prognosis of positive cases of cancer occurrence but decreases the general probability of correct diagnosis. To make use of properties of both approaches we coupled the previously described systems into one system extended by additional networks for cancer diagnosis using only blood parameter data for training. The structure of the whole system is presented in Figure 1-Layer 3. The whole system consists of feed forward and pattern recognition networks usinf as inputs: the vector of tumor markers $\boldsymbol{C}$, the vector of tumor marker with estimated missing values $C_{estimated}$, the combined vector of tumor marker and blood parameter values $BPC = \boldsymbol{(C,P)}$ and the vector $\boldsymbol{P}$ of blood parameters only. The outputs of all networks (prediction of

cancer occurrence) are used as input for the final feed forward network, which calculates the diagnosis.

5 Results

We compare the results of prediction of cancer diagnosis between the system using only tumor marker information and the full system using information coming from tumor marker and blood values. The test data concern breast cancer diagnosis based on four previously described tumor markers and 27 standard blood parameters. The confusion matrices (see Table 4) show results for both systems using the same test data. It can be seen that the combined blood parameter and tumor marker system:

- increases the probability of correct cancer diagnosis (true positives)
- decreases the probability of incorrect diagnosis of positive cancer appearance (false negatives) but also
- increases the probability of incorrect diagnosis of negative cancer occurrence (false positives)

Generally, the introduction of blood parameters makes the system more pessimistic in respect to cancer prognosis: It predicis a positive diagnosis although a cancer disease does not actually exist.

Table 4. Confusion matrix of breast tumor prediction based on C124, C153, C199 and CEA marker group without and with use of blood parameters in neural networks

Neural Networks	P(1/1)	P(1/0)	P(0/0)	P(0/1)	$P_{correct}$
Neural networks system without missing value estimation (vector $\boldsymbol{C}$ as input)	16,4	9,3	58	16,3	74,4
Neural networks system with blood parameters and vector $\boldsymbol{C}$ as input	26,4	17,4	49,9	6,3	76,4

References

1. Jacak W., Pröll K., Data Driven Tumor Marker Prediction System, EMSS 2010, Fes, Marokko, 2010, pp. 1-6
2. Jacak W., Pröll K., Neural Network Based Tumor Marker Prediction, BroadCom 2010, Malaga, Spain, 2010, pp. 1-6
3. Winkler S. M., et al., Feature Selection in the Analysis of Tumor Marker Data Using Evoutionary Algorithms, EMSS 2010, Fes, Marokko, 2010, pp. 1-6
4. Djavan, et al., Novel Artificial Neural Network for Early Detection of Prostate Cancer. Journal of Clinical Oncology, 2002, Vol 20, No 4, 921-929
5. Harrison, et al., ANN models for prediction of acute coronary syndromes using clinical data from the time of presentation. Ann Emerg Med.; 2005, 46(5):431-9.