DATA DRIVEN TUMOR MARKER PREDICTION SYSTEM

Witold Jacak[(a)], Karin Proell[(b)], Herbert Stekel[(c)]

[(a)]Department of Software Engineering
Upper Austria University of Applied Sciences Hagenberg, Softwarepark 11, Austria
[(b)]Department of Medical Informatics and Bioinformatics
Upper Austria University of Applied Sciences Hagenberg, Softwarepark 11, Austria
[(c)] Institute of Laboratory Medicine,
General Hospital Linz, Austria

[(a)] Witold.Jacak@fh-hagenberg.at [(b)] Karin.Proell@fh-hagenberg.at [(c)] Herbert.Stekel@akh.linz.at

ABSTRACT
In this paper a system for the prediction of tumor marker values based on standard blood is presented. Several neural networks are used to learn from blood examination measurements and predict tumor markers in case these values are missing. In a post processing step the predicted values are evaluated in a fuzzy logic like style against different hypotheses and the best hypothesis is used to optimize the predicted values and its plausibility. These predicted values can then be used as input for a second system to support decision making in cancer diagnosis. A variety of experiments with tumor marker C153 show that we can get a prediction accuracy of more than 90%. Our experiments are based on hundreds of samples of up to 27 different features (blood parameters) per vector. We try to predict distinct values, classes of values and a combination of classes and values for specific marker types.

Keywords: neural network, tumor marker prediction, decision support system

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. Examples of tumor markers include CA 125 (in ovarian cancer), CA 153 (in breast cancer), CEA (in ovarian, lung, breast, pancreas, and gastrointestinal tract cancers), and PSA (in prostate cancer). Although an abnormal tumor marker level may suggest cancer, tumor markers are not sensitive or specific enough for a reliable cancer diagnosis. But abnormally altered tumor marker values indicate a need for further medical examination.

During blood examination only a few tumor marker values are tested and for this reason the usage of such incomplete data for cancer diagnosis support needs estimation of missing marker values. Neural networks are proven tools for prediction tasks on medical data (Penny and Frost 1996). For example neural networks were applied to differentiate benign from malignant breast conditions base on blood parameters (Astion et Wilding 1992), for diagnosis of different types of liver disease (Reibnegger et al. 1991), for early detection of prostate cancer (Djavan et al. 2002; Matsui et al. 2004), for studies on blood plasma (Liparini et al. 2005) or for prediction of acute coronary syndromes (Harrison et al. 2005).

In this work we present a novel heterogeneous neural network based system that can be used for tumor marker value prediction. We use an n-dimensional vector of blood parameter values of a several hundred patients as input and train three neural networks in parallel to predict distinct values, classes of values and a combination of classes and values for specific marker types. In a post-processing step the outputs of all networks are adjusted by a fuzzy logic like decision system to obtain the most possible prediction. Unfortunately neural networks are unable to work properly with incomplete data; missing values however are a common problem in medical datasets. It may be that a specific medical procedure was not considered necessary in a particular case or that the procedure was taken in a different laboratory with the values not available in the patient record, or that the measurement was taken but not recorded due to time constraints.

In our system we use two different approaches for dealing with missing values. We reduce the input blood parameters to obtain reasonable complete sample sets. In order to train the networks with complete input blood parameters with impute on missing data values a penalty value during normalizing process.

2. GENERAL CANCER DIAGNOSIS SUPPORT SYSTEM

We focus our considerations on the design of a complex decision support system for early recognition of possibility of cancerous diseases. The system consists

* The work described in this paper was done within the Josef Ressel research center for heuristic optimization sponsored by the Austrian Research Promotion Agency (FFG).

of two components; a Tumor Marker Prediction System and a Diagnosis Support System (see Fig.1). Both systems use several heterogeneous artificial neural networks in parallel. The Tumor Marker Prediction System is in support of the Diagnosis Support System, which uses input data coming from the vector of tumor marker values $\mathbf{C} = (C_1, \ldots, C_m)$ and calculates the possibility of presence of a cancerous disease in general and the possibility of a specific tumor type in particular (tumor types are coded according to ICD 10 system).
The output values are evaluated against different hypotheses and the best hypothesis is used to optimize the predicted diagnosis and its plausibility.

An important issue for the Diagnosis Support System is data incompleteness. We need thousands of vectors of tumor markers for training and evaluation of the neural networks. Those vectors do not only contain distinct values of various marker types but also ranges of marker values, so called classes (e.g. one class for normal values, one for extreme normal values, one for beyond normal but plausible values and one for extreme values). Many of the available marker vectors consist of just a few measurements and cannot be used as training data for neural networks without further processing.

One approach to overcome this problem is to restrict the analysis only to vectors with complete data but this leads to very small sample sets. Another option is to extend the number of input values to all parameters of the blood examination, thus including also non-marker values, using a whole blood parameter vector $\boldsymbol{p} = (p_1, \ldots, p_n)$ as input. Frequently also this vector is incomplete too. For this reason the Tumor Marker Prediction System is connected ahead the Diagnosis Support System, which uses complete or partial complete blood parameter vectors $\boldsymbol{p}$ of patients to train a couple of neural networks for estimating values or classes of values for tumor markers. The output values are evaluated against different hypotheses and the best hypothesis is used to optimize the predicted value. Additionally a possibility value for estimated marker value is calculated. This Marker Value Prediction System could be also be used as a stand-alone system for a rapid estimation of marker values.

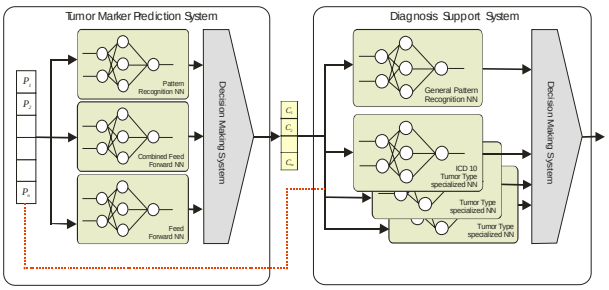

Figure 1: Architecture of Data Driven Cancer Diagnosis Support System

2.1. System for prediction of tumor marker values

Typically, the blood examination results in maximal 27 values of 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 and about 35 tumor markers such AFP, C125, C153, C199, C724, CEA, CYFRA, NSE, PSA, S100, SCC, TPS etc. Many patient records have missing data values as many measurements did not happen being not of interest in a specific case. Different labs also may consider slightly different marker levels to be normal or abnormal. This can depend on a number of factors, including a person's age and gender, which test kit the lab uses, and how the test is done. For each parameter and marker we use the following reference ranges. (See an example in Table 1)

Table 1: Example of blood parameter ranges

CODE	Sex	Normal Lower Bound	Normal Upper Bound	Extrem Norm. Upper Bound	Over Norm. Plausible Upper Bound	Type	Unit	Age LB	Age UP
AFP	M	0	5,8	28	99	AFP (CL)	IU/ml	0	199
AFP	W	0	5,8	28	99	AFP (CL)	IU/ml	0	199
ALT	M	5	45	135	247,5	ALT (GPT)	U/l	0	199
ALT	W	5	34	102	187	ALT (GPT)	U/l	0	199
AST	M	5	35	105	192,5	AST (GOT)	U/l	0	199
AST	W	5	31	93	170,5	AST (GOT)	U/l	0	199
BSG1	M	3	8	15	55	Sinking 1h	Mm	1	199
BSG1	W	6	11	20	55	Sinking 1h	Mm	1	199
BUN	M	5	18	50	165	BUN	Mg/dl	2	16
BUN	W	5	18	50	165	BUN	Mg/dl	2	16
BUN	M	6	20	50	165	BUN	Mg/dl	19	199
BUN	W	6	20	50	165	BUN	Mg/dl	19	199

We divide the value range of marker $\mathbf{C}$ and blood parameter $\boldsymbol{p}$ into k non-overlapping intervals, called classes. In our case study we define four classes ($k = 4$): *Class 1* includes all values less than the Normal Value of marker or blood parameter, *Class 2* includes all values between Normal Value and Extreme Normal Value of marker or blood parameter, *Class 3* includes values between Extreme Normal Value and Plausible Value of marker or blood parameter and *Class 4* include all values greater than the Plausible Value.

For each class i of marker values we calculate the average value μ_i and the standard deviation σ_i. For example the respective values of marker C 153 calculated from patient data are presented in Table 2.

Table 2: C153 tumor marker parameters

Code	*Class*	μ	σ	*Min*	*Max*	*dmax*
C153	1	15,54	5,02	2	25	100
C153	2	33,59	6,73	26	50	100
C153	3	68,48	13,78	56	100	100
C153	4	162,20	321,22	101	10000	100

These classes and their limits are used for the normalizing process of parameter and marker values.

2.1.1. Architecture of marker value prediction system

The system consists of three heterogeneous parallel coupled artificial neural networks and a decision-making system based on aggregation rules.

Normalizing: The input and output values for training and testing of each network 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. Such a normalizing process guarantees that all values are mapped to interval [0, 1].

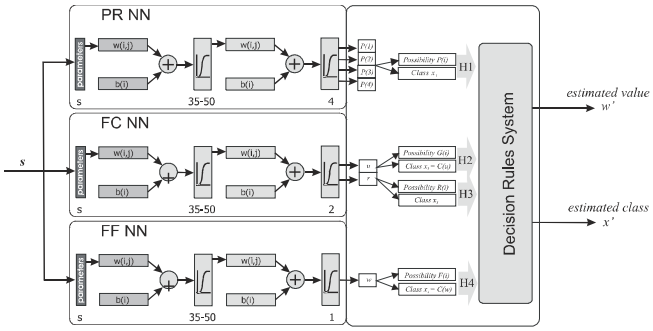

Figure 2: Neural networks based tumor marker value prediction system

Prediction System Structure: The general marker value estimation system contains of three neural networks (see Figure 2).

- Feed forward neural network (FF) with ***p*** inputs (normalized values of blood parameter vectors ***p***) and one output, normalized values of marker *C*

- Pattern recognition neural network (PR) with ***p*** inputs (normalized values of blood parameter vectors ***p***) and *k* outputs, *k*-dimensional binary vector coding classes of marker *C*

- Combined feed forward neural network (FC) with ***p*** inputs (normalized values of blood parameter vectors ***p***) and two outputs: normalized values of marker *C* (as in network FF), and normalized classes of marker *C* as:

$$NormClass^{j}(C) = j/k, \quad \text{for} = 1,...,k \qquad (1)$$

All neural networks have one hidden layer and tan-sigmoid or log-sigmoid transfer function. The output values of neural networks belong usually to the interval [0, 1].

For a given parameter vector ***p*** the three networks calculate different output data.

- The pattern recognition neural network PR produces the *k*-value vector $(P(i) \mid i = 1,..,k)$, where $P(i)$ describes the possibility (in sense of fuzzy logic) of class *i* of *C* marker connected to input parameter vector ***p***. The supposed class of marker *C* is

$$x_1 = arg(max\{P(i) | i=1,...,k\}). \qquad (2)$$

- The two output feed forward neural network FC, combined value and class training, generates two values, the value *r* interpreted as possibility of class marker *C* and value *u*, which is predicted normalized value of marker *C*.
  For value output *u* (similarly to the output of FF network) and limits of marker values we can calculate the class $x_2 = Class(u)$ to which belongs the output *u*.
  Separately we calculate the possibility vector $(G(i) \mid i = 1,...,k)$ as indirect distance to average value of each *C* marker class

$$G(i) = 1-(|u-\mu_i|/d_{max}) \text{ for } i = 1,...,k. \qquad (3)$$

  where d_{max} denotes maximal distance between values of marker *C*.
  For *r* value we calculate the possibility vector $(R(i) \mid i = 1,...,k)$ as

$$R(i) = 1-|r-i/k| \text{ for } i = 1,...,k. \qquad (4)$$

  The class of marker *C* that will be suggested is

$$x_3 = arg(max\{R(i)|i=1,...,k\}). \qquad (5)$$

- The feed forward neural network generates a normalized value *w* of marker *C*. Based on value *w* and the limits of each class we can calculate the class $x_4 = Class(w)$ to which the output of FF network belongs.
  Separately we calculate the possibility vector $(F(i) \mid i = 1,...,k)$ as indirect distance to average value of each *C* marker class

$$F(i) = 1-(|w-\mu_i|/d_{max}), \text{ for } i = 1,...,k. \qquad (6)$$

We use of the Neural Network Toolbox™ of MATLAB® for designing, implementing, visualizing, and simulating the neural networks PR, FC and FF.

2.2. Evaluation and post processing method

Based on the calculated estimation of marker values we can establish four hypotheses x_1, x_2, x_3, x_4 for determination of classes. For each hypothesis x_1, x_2, x_3, x_4 the possibility value *P*, *G*, *R*, *F* is calculated too.

These hypotheses should be verified to find the maximal possible prediction. This is done by testing a couple of aggregation functions *V* on each hypothesis possibility values. Those aggregation functions are

similar to aggregation rule in fuzzy logic decision-making system. In this experiment we use four kinds of functions *V*:

Minimum: $V(x_i) = min\{P(x_i)\ G(x_i)\ R(x_i)\ F(x_{ij})\}$ (7)

Product: $V(x_i) = P(x_i)\ G(x_i)\ R(x_i)\ F(x_i)$ (8)

Average: $V(x_i) = (P(x_i) + G(x_i) + R(x_i) + F(x_i))/4$ (9)

Count: $V(x_i) = Count\_of(x_i)$ (10)

Count is the number of identical hypothesis.

The maximal value of aggregation function determines the new predicted class. i.e.

$$x_{new} = arg\ (max\{V(xi) \mid i=1,..,4\}) \quad (11)$$

where $V(x_i) = V(P(x_i), G(x_i), R(x_i), F(x_i))$ is the evaluation function of arguments $P(x_i)$, $G(x_i)$, $R(x_i)$ and $F(x_i)$. This kind of evaluation method leads to four decision composition rules, namely *MaxMin*, *MaxProd*, *MaxAvg* and *MaxCount* known in fuzzy decision systems.

Estimation of predicted marker value: Based on such determined new class the estimation of marker value is performed. If the evaluation function *V* used in decision composition rule has a value greater than τ then the new estimated value of marker is equal to the average value of *w* and *u*. i.e.

$$w_{new} = (w + u)/2 \quad if\ V(x) > \tau \quad (12)$$
$$w_{new} = (w + u + \mu_{xnew})/3 \quad other$$

3. CASE STUDY: C 153 TUMOR MARKER

3.1. Training and Test Setup

We have taken the complete data set with 20 blood parameters from patient data: The input vector ***p*** contains complete data of following blood parameters ***p*** = (HB, WBC, HKT, MCV, RBC, PLT, KREA, BUN, GT37, ALT, AST, TBIL, CRP, LD37, HS, CNEA, CMOA, CLYA, CEOA, CBAA). We use 4427 samples as Learning Pattern Set for the neural networks system and 491 independent samples as Test Set. The datasets based on the whole data set the Pearson correlation coefficient between tumor markers and blood parameter is calculated. This correlation matrix in % is shown in Table 3:

Table 3: Correlation between tumor markers and blood parameters

Marker/ Parameter	AFP	C125	C153	C199	C724	CEA	CYFS	FPSA	NSE	PSA	PSAQ	S100	SCC	TPS
ALT	16	-10	25	19		10			31			-10		49
AST	33	13	40	27	15	27	11		49			8		36
BSG1		22	13	9	50	13	34		23	15		119	25	15
BUN		17					25	18	15		19	74	10	8
CBAA			-10									-11		
CEOA	-11		-13		-11									9
CH37	-28	-39		-23	-16	-9	-35	-11	-21				-12	-13
CHOL		-20			14		-19					-39		
CLYA	-19	-36	-20	-12	-20	-16	-15		-14	-13	-9	-35	-14	
CMOA		18			19	10							8	
CNEA		25	19		11	12	17	10	11	11		21	14	
CRP	18	32	27	17	23	23	20		41	21			11	
FE	11	-18	-29		-24	-12	-16						-11	-27
FER	15	31	31	32			35		59		14			-11
GT37	22	15	37	39	50	30			43					30
HB		-35	-38	-25		-19	-18			-15		-12	-17	
HDL	-11			-16								-151		24
HKT		-32	-36	-25		-16	-13		-8	-14			-16	
HS	-15	19		-12	-15						9			9
KREA		15	8	-9	9		14	17			21	64	10	
LD37	19	34	45	23	29	35	40		69	14		70		
MCV	17			10									-12	
PLT	-15	10	16		12		16				-11	17	12	12
RBC		-30	-31	-28		-16	-9			-12			-11	
TBIL	18	13		19	-16	12			21		9			
TF		-55	-29	-26	-28	8	-32	-20	-25				-12	34
WBC							22	11	17			24	14	

Data normalization: All data are normalized (on values [0, 1]) based on 3 classes (Normal Value, Extreme Normal Value, and Over Normal but Plausible Value). Values over Plausible value are the normalized with value 1. Data taken for learning and test process include the all four classes.

3.2. Experiments and Results

For establishing the number of neurons in a hidden layer of feed-forward networks we performed small batch set training of networks using different numbers of neurons.

The result is presented in Fig. 3.

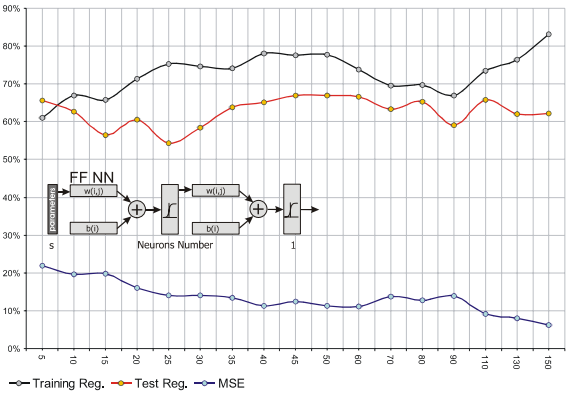

Figure 3: Test regression of neural networks with different number of neurons

The empirical test shows that best performing are networks with 40-60 neurons of hidden layer. We chose neural networks having one hidden layer with 50 neurons and tan-sigmoid activation functions. These are the neural networks settings used:

- Feed forward neural network (FF) with 20 inputs (normalized values of parameter vectors ***p***) and one output (normalized values of C153 marker)
- Pattern recognition neural network (PR) with 20 inputs (normalized values of parameter vectors ***p***) and four outputs (four dimensional binary vectors coding classes of C153 marker)
- Feed forward neural network (FC) with 20 inputs (normalized values of parameter vectors ***p***) and two outputs (normalized values of C153 marker and normalized classes of C153 with: class 1= 0,25; class 2= 0,5; class 3= 0,75; class 4=1,0)

All three neural networks were trained with Levenberg-Marquardt algorithm and a validation failure factor 6. The test outputs of all network is post processed by aggregation rules based on the decision system and finally compared with original values and classes of C153 tumor marker from test set.

The regression functions between the outputs of these three networks, post processed final estimation

and test C153 values are presented in Figure 4. It can be observed that regression of the rules based final estimation of C153 value is greater (R =0,73) than the individual estimation of separate networks (R= 0,65, R = 0,68, and R = 0,65 for PR, FC, and FF network respectively). The blue bounded area marks fatal mismatching i.e. originally large values of C153 marker and small-predicted values.

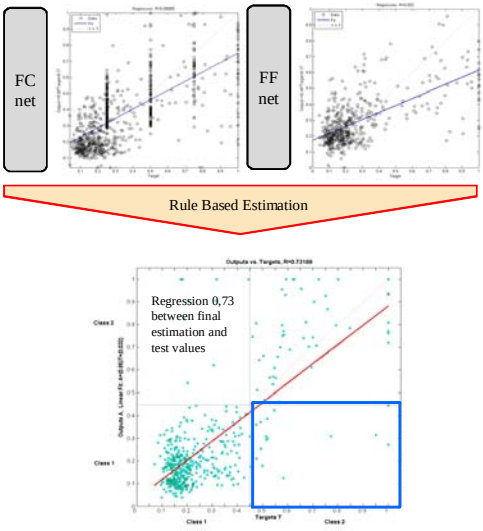

Figure 4: Results of rule based estimation system for tumor marker C153

Based on the predicted values of tumor marker C153 we calculate the matching and mismatching ratios for each class of marker. Matching and mismatching of classes between the test data of marker C153 classes and the predicted classes of different networks are presented as confusion matrixes (see Fig. 5). After post processing, we obtain 72 % matching of four classes, whereas the original networks range between 59% and 66% matching cases. The ratio of fatal mismatching is 2,9 %.

Class Net / Class Target	1	2	3	4	
1	54,8%	9,2%	0,2%	0,0%	
2	11,2%	10,0%	1,6%	0,6%	
3	0,6%	1,4%	5,9%	1,4%	
4	0,4%	0,4%	1,2%	1,0%	
Gesamt System MaxAvg	2,9%				71,7%

Confusion matrix of final estimation

Figure 5: Confusion matrix of results of Rule Based Estimation System for tumor marker C153

Additionally we have performed an experiment with a reduced number of training classes of marker C153. We merge the class 1 and class 2 of marker C153 into a new class I and class 3 and class 4 into a new class II. That means that all values of tumor marker C153 less than Extreme Normal Value of C153 determine class I and values greater than Extreme Normal Value determine class II. Normalization of blood parameters remains unchanged. Full training and test of networks was performed on input data modified in this way. The test results for those two classes are shown in separate confusion matrix tables (see Fig. 6).

Confusion matrix PR NN

Class Net / Class Target	Class I	Class II	
Class I	82,1%	8,9%	
Class II	2,4%	6,5%	
Gesamt PR Net			88,6%

Rules Based Estimation

Confusion matrix final estimation

Class Net / Class Target	Class I	Class II	
Class I	84,3%	3,3%	
Class II	2,0%	10,4%	
Gesamt System MaxAvg			94,7%

Figure 6: Confusion matrices before and after application of rules based estimation

We obtain a quite good ratio (94,7 %) of matching and the ratio of fatal mismatching is reduced to 2 %.

The dependency of quality of estimation on different composition rules is presented in Table 4. It is shown that the *MaxAvg* rule produces the best results.

Table 4: Results after application of composition rules

Rule	Confusion PR Net	Confusion FF Net	Confusion FC Value Net	Confusion FC Class Net	Confusion New Value	Confusion Two Class (Class based)	Confusion Two Class (Value based)
MaxCount	68,2	59,5	66,1	63,5	67,6	93,7	94,5
MaxAvg	68,2	59,5	66,1	63,5	**71,7**	**94,7**	**94,7**
MaxMin	68,2	59,5	66,1	63,5	68,2	89,8	94,3
MaxProd	68,2	59,5	66,1	63,5	68,0	90,2	94,5

3.2.1. Experiment with missing blood parameters values

Additionally, we have performed an experiment using vectors containing numbers of measured values from 27 to 15 (maximal 12 missing values allowed), as the number of vectors containing all 27 parameters is very small (about 160 samples). During normalization we replace the missing values with the value -1 thus resulting in input vectors containing all 27 blood parameters. We have obtained 6191 samples as Learning Pattern Set for the neural networks system and 618 independent samples as Test Set. All three neural networks were trained by Levenberg-Marquardt algorithm using a validation failure value of 6. The test outputs of all networks are post processed by the decision system based on aggregation rules and finally compared with original values and classes of C153 tumor marker from test set.

Confusion matrixes of final class estimation (Table 5) show that the final estimation method works well too with input vectors containing missing values. After post processing, we achieve 67 % matching of four classes, whereas the original networks gets between 53% and

61% matching cases. The ratio of fatal mismatching increases to 5,8 %.

Table 5: Confusion matrixes of final class estimation using input vectors with missing values and four classes

Class Net / Class Target	1	2	3	4	
1	47,1%	13,3%	0,2%	0,2%	
2	8,9%	15,0%	2,3%	0,0%	
3	0,3%	4,5%	3,7%	1,3%	
4	0,3%	0,6%	1,1%	1,1%	
Gesamt System MaxAvg	5,8%				**67,0%**

The test results for two classes are shown in a separate confusion matrix (see Table 6). We obtain in this case a good ratio (90,5 %) of matching and the ratio of fatal mismatching is reduced to 4,7 %.

Table 6: Confusion matrixes of final class estimation using input vectors with missing values and two classes

Class Net / Class Target	**Class I**	**Class II**	
Class I	82,0%	4,9%	
Class II	4,7%	8,4%	
Gesamt MaxAvg			**90,5%**

The regression functions between post processed final estimation and test C153 values are presented in Fig. 7.

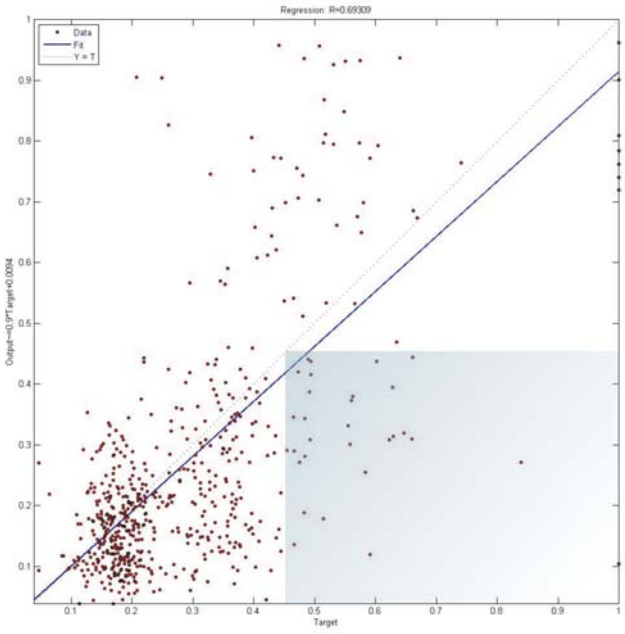

Figure 7: Regression functions between post processed final estimation and test C153 values

4. FINAL REMARKS

In this paper we focused our presentation on experiments predicting the marker C153. Similar results were obtained for marker C125, regression value 0,7865, for marker C199, regression value 0,7245 and for marker CEA, regression value 0,7225. In the next experiments we will continue testing further marker types depending on the availability of sufficient patient data. It can be expected that not all markers can be predicted with similar performance as C153.

In a further step the system will be extended by predicting combinations of markers as output of neural networks. Short tests show that such combinations can increase the quality of prediction. Moreover it is seems to be more effective to create different classes of markers as used in the recent experiment based on self-organizing maps methods. This prediction system will be an important component to the whole data driven cancer diagnosis support system.

ACKNOWLEDGMENTS

Special thanks goes to Prim. Herbert Stekel, M.D, Head of Institute of Laboratory Medicine, General Hospital Linz, Austria for providing thousands of samples of anonymous blood examination data for our experiments.

REFERENCES

Astion, M.L., Wilding P., 1992, Application of neural networks to the interpretation of laboratory data in cancer diagnosis. *Clinical Chemistry*, Vol 38, 34-38.

Djavan, B., Remzi, M., Zlotta, A., Seitz, C., Snow, P., Marberger, M., 2002, Novel Artificial Neural Network for Early Detection of Prostate Cancer. *Journal of Clinical Oncology*, Vol 20, No 4, 921-929

Harrison, R.F., Kennedy, R.L., 2005, Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. *Ann Emerg Med*.; 46(5):431-9.

Liparini, A., Carvalho, S., Belchior, J.C., 2005, Analysis of the applicability of artificial neural networks for studying blood plasma: determination of magnesium on concentration as a case study. *Clin Chem Lab Med*.; 43(9):939-46

Matsui, Y., Utsunomiya, N., Ichioka, K., Ueda, N., Yoshimura, K., Terai, A., Arai, Y., 2004, The use of artificial neural network analysis to improve the predictive accuracy of prostate biopsy in the Japanese population. *Jpn J Clin Oncol*. 34(10):602-7.

Penny, W., Frost, D., 1996, Neural Networks in Clinical Medicine. *Med Decis Making*, Vol 16: 386-398

Reibnegger, G., Weiss, G., Werner-Felmayer, G., Judmaier, G., Wachter, H., 1991, Neural networks as a tool for utilizing laboratory information: Comparison with linear discriminant analysis and with classification and regression trees. *PNAS:* vol. 88 no. 24, 11426-11430