Medical Informatics Europe '97
C. Pappas et al. (Eds.)
IOS Press, 1997

The Knowledge Model of MedFrame/CADIAG-IV

Barbara Sageder, Karl Boegl, Klaus-Peter Adlassnig, Günter Kolousek, Bernhard Trummer

Department of Medical Computer Sciences, University of Vienna, Medical School, Austria
Spitalgasse 23, A-1090 Vienna, Austria
E-Mail: b.s@trulli.imc.akh-wien.ac.at

The medical consultation system MedFrame/CADIAG-IV is a successor of the prior CADIAG projects. It is the result of a complete redesign to account for today's demands on state-of-the-art software.
Its knowledge representation and inference process are based on fuzzy set theory and fuzzy logic. Fuzzy sets are used for conversions from measured numeric values and observational data into symbolic ones. Medical relationships between findings, diseases, and therapies, the rules, are represented by fuzzy relations, that express positive or negative associations. Findings, diseases, and therapies are organised in hierarchies.

1. Introduction

MedFrame/CADIAG-IV is a rule- and frame-based diagnostic and therapeutic consultation system that is under development at the Department of Medical Computer Sciences of the University of Vienna Medical School. It is the latest successor of the CADIAG projects that started with a system based on Boolean logic in 1968. CADIAG-2 [1] and its successors were based on fuzzy set theory and fuzzy logic to handle inherent vagueness and uncertainty of medical knowledge.

MedFrame/CADIAG-IV consists of the following components [4]:

- patient management and general administration
- database
- knowledge base
- inference engine
- explanation tool
- user interface

In this paper we introduce MedFrame/CADIAG-IV. We will especially focus on its knowledge representation, which is based on fuzzy set theory.

2. New Features of MedFrame/CADIAG-IV

The target of MedFrame/CADIAG-IV has been the adjustment to today's demands on a modern computer environment. Accordingly, it is designed as a multi-user, platform-independent client/server system that is implemented by use of flexible object-oriented

programming and modelling techniques. Furthermore, it is equipped with a modern graphical user interface with multi-language capability and comprehensive help-functions.

A problem that has to be solved is that in a multi-user system users' access rights to data must be managed. In MedFrame/CADIAG-IV this is done in a Unix-like manner. To unify data representation, a multi-user system also requires a standardised terminology. In our system this is based on SNOMED* International [2]. Also important for a modern system is platform-independence, which we achieved by using Smalltalk-VisualWorks.

Another new feature is the possibility of communication with other systems via computer networks and internet. So data can be exchanged between various different medical computer systems, either within the hospital or world-wide. Likewise a user can access MedFrame/CADIAG-IV via internet. Here the problem of data protection has to be solved to prevent abuse of the information about patients.

3. Knowledge Representation

The interrelations between the major components of knowledge representation and inference are depicted in Fig. 1. These components are described in more detail below.

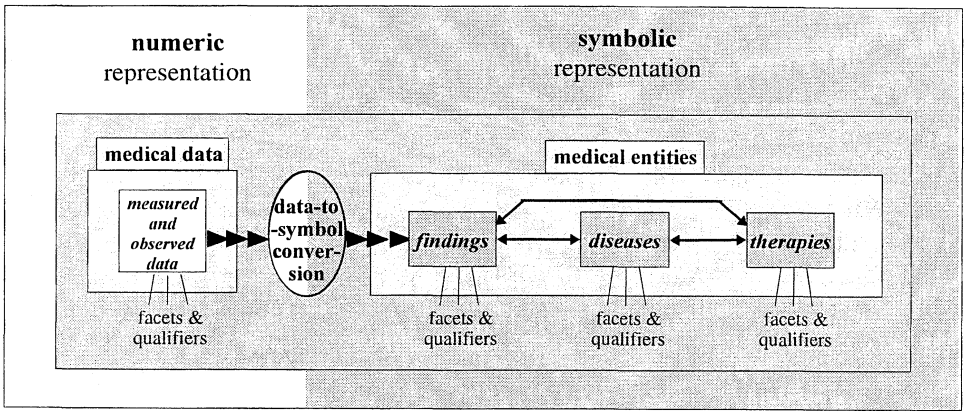


Fig.1 General structure of the knowledge model of MedFrame/CADIAG-IV.

3.1 Medical Concepts

Medical concepts are differentiated into *medical data* and *medical entities*. Medical entities are further divided into *findings*, *diseases*, and *therapies*.

Medical data mean exact values whereas medical entities mean symbolic concepts. So 'temperature: 39°' would be a medical data, 'high fever' a medical entity. Medical knowledge is normally formulated using symbols, so the inference process of MedFrame/CADIAG-IV is also based on them.

3.2 Data-to-Symbol Conversion

In many cases the input data for the inference process already are medical entities, if not, a conversion, the *data-to-symbol conversion*, has to take place. The *data-to-symbol*

* Standardized Nomenclature of Human und Veterinary Medicine

conversion is based on fuzzy sets. Possible fuzzy sets for the conversion of the body temperature into symbolic concepts are depicted in Fig.2.

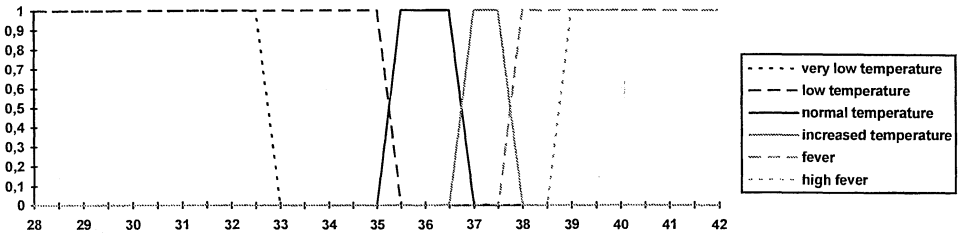


Fig.2 Fuzzy sets for the data-to-symbol conversion of body temperature values.

Fuzzy sets were first introduced by Zadeh in 1965 [9]. Each member of a fuzzy set is associated with a certain degree of membership. So a member does not simply belong to a fuzzy set or not, as it is for a normal set, but belongs to it to a certain degree. Additionally, every value can belong to different fuzzy sets with different degrees of membership.

Type 2 fuzzy sets are fuzzy sets whose degrees of membership are fuzzy sets themselves. So even more complex situations can be expressed vaguely. As an example, type 2 fuzzy sets are used to describe temporal courses. The degrees of membership here are not fixed values but functions changing over time. An example for a 'normal glucose tolerance test' is depicted in Fig.3

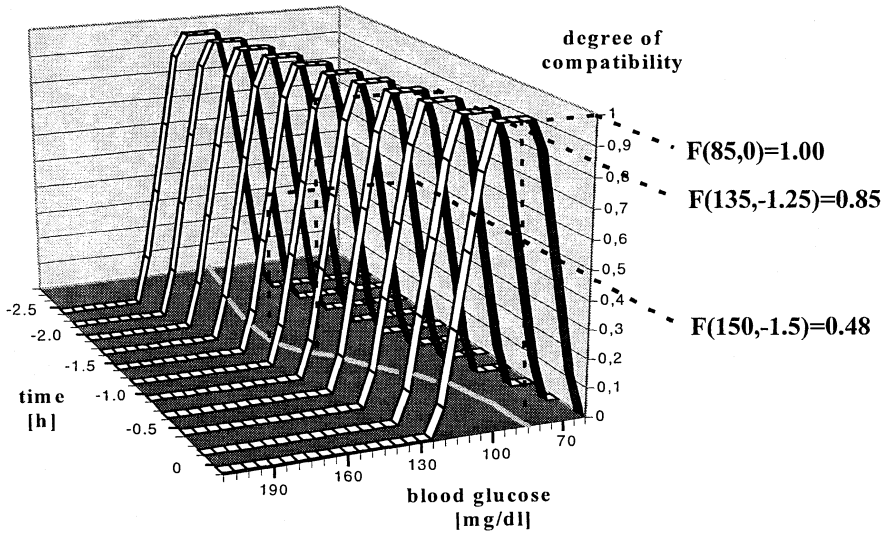


Fig.3 Assessment of a temporal course of blood glucose level by a type 2 fuzzy set

3.3 Rules

Relationships between medical entities are expressed by *rules*, which are fuzzy relations. Fuzzy relations are fuzzy sets of the Cartesian product of two or more sets.

Rules describe positive and negative associations influencing the positive or negative

truth values of the medical entity on the right side of the rule respectively. Positive and negative truth values are collected separately. The positive truth value holds evidence *for* the diagnosis, the negative one *against* it. Both take values between 0 and 1. A value of 0 equals false, 1 equals true. Special cases that have to be handled accordingly are:

- positive value = 1; negative value < 1 - confirmed diagnosis
- positive value < 1; negative value = 1 - excluded diagnosis
- positive value = 0; negative value > 0 - excluded diagnoses
- positive value = 1; negative value = 1 - inconsistency in the knowledge base

Several compositional rules of fuzzy inference are the basic means of doing inferences [1]. At present, studies are under way to select those compositions most adequate (e.g. max-min compositions, sum-min compositions). An evaluation of the inference strategy in CADIAG-2 was done in [2].

3.4 Medical Terminology

All *findings, diseases, and therapies* are organised in hierarchies [7] that group related items together. The structure of these hierarchies is based on ICD-10* [8] and SNOMED [2]. An example for a part of such a hierarchy of diseases, that is based on ICD-10, is depicted in Fig. 4.

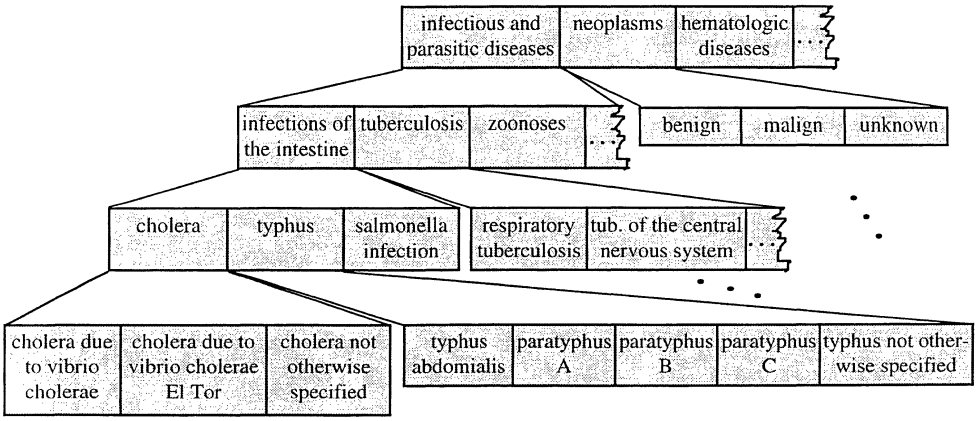


Fig.4 Example of a hierarchy of diseases.

Medical entities can also be described by a more-dimensional classification. In MedFrame/CADIAG-IV *facet*-hierarchies based on SNOMED are used for this purpose. Whenever the combination of *facets* is not exact enough, *qualifiers* (like: left, right, big, little, short ...) complete the description.

4. Discussion

To represent medical knowledge correctly, we must partly rebuild the cognitive process of a physician concluding a specific diagnosis.

Fuzzy sets and fuzzy relations are a very effective and comprehensible means for the

* International Classification of Diseases

representation of vagueness. Prior tests have shown that most physicians can cope with them after some learning period.

Normally a physician does not define a connection between a finding and a diagnosis directly. There are groups of diseases that get evidence from a special finding in the same way. For example the finding 'fever' does not really lead to some special infectious disease but rather to the group of all infectious diseases. So the use of hierarchies of diseases (Fig.4) helps to define realistic rules that are comprehensible for physicians.

A separation of positive and negative truth values is important to compute excluded diseases, which normally result from a negative rule with strength of implication 1. It is easier to have the two values than to handle all possible cases with one value.

Negative evidences as mentioned above can often be converted into positive ones when we again take into account how the physician reasons. When he says that a finding does rather not lead to some disease, he does not mean, that it leads to *all* other existing diseases. What he really means is that it supports a group of diseases *complementary* to this disease in that specific context. Unfortunately not every negative evidence can be converted into positive ones in this manner. For example the negative rule: 'male sex excludes a gynaecological disease' cannot be expressed positively because there is no group 'non-gynaecological disease'.

5. Conclusion

In this paper we introduced the medical consultation system MedFrame/CADIAG-IV. We put special emphasis on the knowledge representation techniques. We discussed the basics of fuzzy set theory as well as the use in the knowledge representation and inference process of MedFrame/CADIAG-IV.

We still have to evaluate the new knowledge model, but first tests have produced strongly acceptable results.

6. Acknowledgement and References

The research on MedFrame/CADIAG-IV was supported by the 'Jubiläumsfond der Österreichischen Nationalbank, Projektnummer 5433'.

[1] Adlassnig K.-P., Fuzzy Set Theory in Medical Diagnosis, IEEE Transactions on Systems, Man, and Cybernetics 16 (1986) 260-265

[2] Côté R.A., Rothwell D.J., Palotay J.L., Beckett R.S., and Brochu L., The Systematized Nomenclature of Human and Veterinary Medicine - SNOMED International, American College of Pathologists, 1993

[3] Daniel M., Hájek P., Nguyen P.H., CADIAG-2 and MYCIN-like systems, Artificial Intelligence in Medicine 9 (1997) 241-259

[4] Kolousek G., Major Design Elements of Cadiag-IV 2.0, Technical Report MES-1996, Department of Medical Computer Sciences, University of Vienna Medical School, 1996

[5] Leitich H., Anforderungen an ein Wissenserwerbssystem für das medizinische Expertensystem CADIAG-IV, Master Thesis, Technical University of Vienna, 1995

[6] Leitich H., Boegl K., Kolousek G., Rotenfluh TE. and Adlassnig K.-P.: A fuzzy model of data interpretation for the medical expert system MedFrame/CADIAG-IV. In: Trappl R (Ed.): Cybernetics and Systems '96, Vol 1. Austrian Society for Cybernetic Studies, 1996

[7] Smutny E., Terminologie für das medizinische Expertensystem CADIAG-IV, Master Thesis, Technical University of Vienna, 1996

[8] WHO, ICD-10 Internationale statistische Klassifikation der Krankheiten und verwandte Gesundheitsprobleme 10.Revision, Ed. Deutsches Institut für medizinische Dokumentation und Information, 1994

[9] Zadeh L.A., Fuzzy Sets, Information and Control 8 (1965) 338-353

Medical Informatics Europe '97
C. Pappas et al. (Eds.)
IOS Press, 1997

Studies of Responsiveness in the R-EGFR and Rcerb-B2 Oncogenes Spaces: a Customized Application for Thyroid Lesions

Arijan Šiška[a] B. Sc., Ankica Babić[b] Ph. D., Nikola Pavešić[a] Ph. D.

a) Faculty of Electrical Engineering, Tržaška 25, Ljubljana
b) Department of Biomedical Engineering, Medical Informatics, Linköping University, Sweden

Abstract

Among verity of diagnostic approaches suitable for clinical analysis of thyroid lesions, the two oncogenes (R-EGFR and Rcerb-B2) are believed to be of discriminative power. In a retrospectively collected patient material we have defined different lesion types (normal tissue, benign and malignant tumours). Those were taken as class definitions in analysis performed to assign discriminating performance. Standard multivariate statistics has not performed satisfactory partly due to the data distribution and partly due to presence of the noise. Therefore we have developed a method for the classification purpose, which was based on principle of minimising generalised classification error. Results of the separation between carcinoma and normal tissue reached accuracy 70%, other classification attempts ended up in poor results. In general, misclassifications could be explained with the data quality (noise) and, when it came to benign lesions, with responsiveness of the oncogenes to tumour tissues.

Introduction

Follow up of patients with thyroid related complications is done with a purpose of controlling benign lesions and disclosing malignant processes. Routinely performed tests include a specific laboratory profile, histological and cytological examination. Exposition of the oncogenes is expected to be a sophisticated measure reflecting subtle changes in the tissue[1,2,3]. Aim of the present study was to assess this performance by means of statistical and other applicable methods. We approached this problem in a common manner by finding classification rules for the known data sets which could be then hypothetically used to diagnose new cases.
We studied thyroid lesions of 97 patients, 61 (62%) with benign and 37 (38%) with malignant changes.

2D formulation of the classification

Univariate and multivariate statistics were used to explore the patient material with respect to a great number of clinical and diagnostic features. The data distribution suggested nonparametric tests to be performed, and additional efforts to discriminate the groups. Due to the noise in data, most likely caused by time span in which patients were treated, preliminary accuracies were not high. Therefore, we had decided to search for more efficient classifiers, thus we have chosen a 2D representation. That meant that the diagnostic problem was turned into a problem of dividing-partitioning a set of points that belong to different classes in N dimensional space. This partitioning can be done in various ways of which the simplest and most intuitive might be to partition by means of linear functions or hyper planes. We are dividing N dimensional space with N-1 dimensional hyper planes into two subspaces. If we recursively divide subspaces into smaller subspaces, we get a hierarchical division of space with hyper planes into arbitrary number of parts or classes. Such division is called binary space partition. Hierarchical structure of division planes is called BSP tree. Problem can best be viewed in the 2D plane. Points are two dimensional, division planes are 1 dimensional lines. Subsequent generalisation into N dimensional problem is easy and can be derived later. All is a part of a larger class of classification problems some of which hypothesise about measured data distribution functions or pay a lot of attention to special points which reside near the border [4,5,6].

Classification Method

First of all, to define fitness of the proposed division, we have to ask ourselves: "How good does this line divides the plane?". Let us first define the error function E. This is a function that, for a given set of points in the plane, and with each point having defined a class membership, and given proposed dividing line, produces a

real number. That is our measure of error, i.e. criterion of fitness for the given dividing line. The smaller error, the better division. Formally we can define the function E as:

$$E:\{k,n,p_1,p_2,\ldots,p_n,c_1,c_2,\ldots,c_n\} \mapsto e$$

where k and n are linear function coefficients, p_i are points on the plane, c_i are class indexes and e is the error value. Error function defines the behavior of the dividing line. Out of all possible functions we must extract a class of functions that have some sort of meaning. Some attributes of the error function can be intuitively recognized:

- points-samples p_i are equivalent and error function is invariant with respect to these points. If we shuffle the points' indexes, for example, the error should remain the same.
- if we put all points from the first class into the second class, and all points from second class into first class, error should remain the same.
- error function is depended on a signed distance of point p_i from the line: this distance can be defined as Euclidean distance or: $(p_i\text{-}[O,n]).[k,-1]$, which is a principle of linear discrimination described in [5].

Following the above, we developed error function E, that is defined as a sum of errors that single points produce:

$$e = \sum_i f\left(\left(\left(p_i - [0,n]\right).[k,-1]\right) * c_i\right)$$

As we can see from the equation, there is a function f acting as a modifier of the error of a single point. This function is a generalization of the principle of peceptron criterion function [5]. Minimizing this criterion-error function is our goal.

There are only two unknown variables in the equation k and n, that define the dividing line. Our task is to find such parameters that minimize error e.

Method as just described can be viewed as a penalty method. We prescribe the penalty line has to pay if it sets itself on the "wrong side" of the point. When we minimize error function we actually try to enforce restrictions upon line placement.

What we are basically interested in is such an error function, that counts the wrongly classified points-samples. In this case we are minimizing number of miss-classified points. Function f in that case is defined as a step function. This function is not continuous and that fact can pose a real problem to various algorithms for finding a minimum.

If we chose f to be a right linear function we effectively get the method of minimizing perception criterion function [5].

At this point it is probably interesting to notice, that if we put $f(x) = x^2$ we practically get a method of placing a linear regression function through the points. We are looking for a linear function that has minimum square distance to the points. Of curse in this case there is no classification taking place since function f is symmetrical with respect to the y coordinate axis and therefore in a way blind to the point's class.

We can also define different f, with following properties:

- separate left and right side of the y axis in the coordinate system, by being small in the left side and growing large on the right side
- are continuous

Let us take a look at the Table 1.

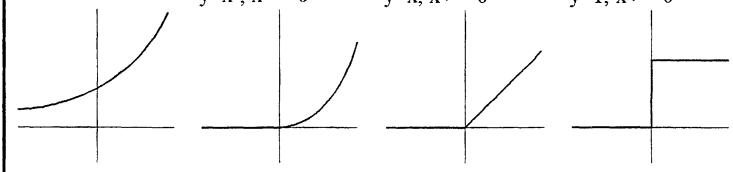
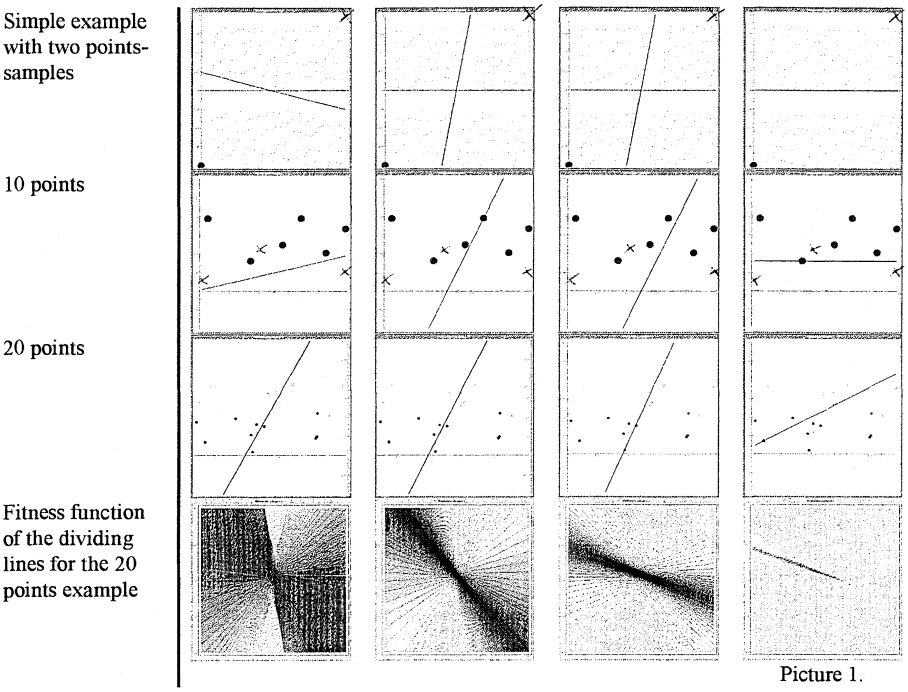
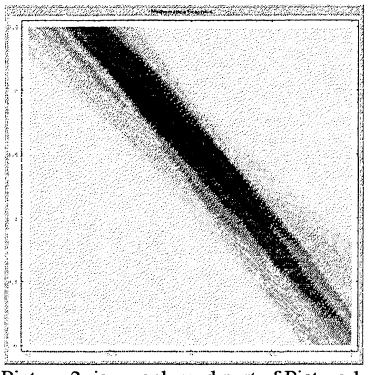
function f	Exponential function	Right Square function	Right Linear function	Step function
definition	$y=e^x$	$y=0,\ x<0$ $y=x^2,\ x>=0$	$y=0,\ x<0$ $y=x,\ x>=0$	$y=0,\ x<0$ $y=1,\ x>=0$
function graph				

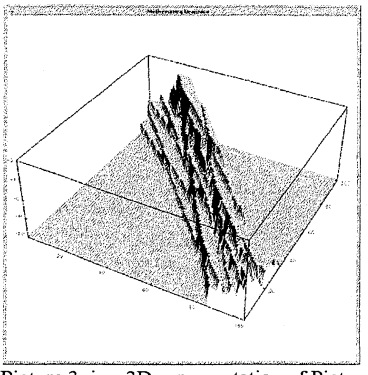
Table 1. A graphical presentation of choices for the f function and its corresponding effect on derivation of minimal error dividing line - classification criterion.



Picture 1.



Picture 2. is an enlarged part of Picture 1



Picture 3. is a 3D representation of Picture 2.

Enlarged part of the Picture 1. and its three dimensional display clearly suggest a difficulty of finding extreme of the error-fitness function with f being a step function. The error function is not continuous which makes it practically impossible to locate a minimum point. Even if the step function f is theoretically our ideal goal, since error function returns a number of miss-classified samples and that being the only relevant number to us, its use is clearly limited, since many algorithms for minimization depend on using derivation or gradient of the error function. This is something we cannot do if error function is not continuous.

As a compromise solution it is suggested to take a right linear function [5]. This would make error function continuous and enable use of gradient methods, but still derivation of the error function is not continuous and that makes it impossible to use second derivation in the minimization process. Still the fitness relief for both right linear function and step function are similar enough, to make us believe that we can get away by using right linear function for function f instead of the real thing.

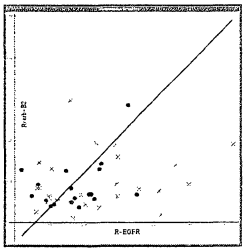
There are many different algorithms which find extreme points or minimum of the given function, and it is certainly not the point of this paper to go into many details regarding these methods. At this point genetic algorithms should be mentioned as a promising direction.

What we have come up with in this test example is a fairly simple method that tries to minimize parameters separately, first k than n, and then the whole scheme repeats. At first steps at which parameters are changed are big, then they gradually get smaller and smaller. Since fitness landscapes in our example are simple, this procedure works well enough. However more research in this part would be required to produce more accurate results faster, or prove there can be no significantly more accurate results.
We hoped that the method described will be less influenced by necessary noise in the data, since it does not necessarily tries to put all the points on the right side of the classification line, but it tries to minimize an error criterion. By introducing noise into data we hoped classification line will not change it's position very much. This robustness for noise was one of our goals.
All the programming was done in the Mathematica [7],[8]. programming environment. It is not very fast, but it is flexible enough for us the make that trade-off. This is especially important when trying out lot's of different algorithms.

Results

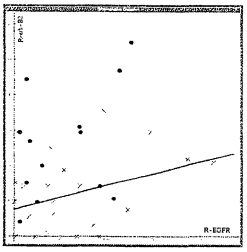
We applied our procedures to the real data. As a function f we used right linear function. Points are plotted in the plane, where X axis represents R-EGFR oncogene measurement and Y axis represents Rcerb-B2 oncogene measurement. An example of the classification is given in the pictures 4 and 5.
Points were separated into two classes by the following criteria:

class 1 (bright crosses) under the line: Benign tumours;
class 2 (dark points) above the line: Normal tissue;
dividing line: y=1.738*x-50;
classification error: 22 miss-classified samples out of 46



Picture 4.

class 1 (bright crosses) under the line: Malignant tumours;
class 2 (dark points) above the line: Normal tissue;
dividing line: y=0.137*x+15;
classification error: 12/39



Picture 5.

We evaluated our method with standard linear regression tests. These tests helped us find any relationship between the class of a point and its corresponding R-EGFR and Rcerb-B2 parameter.
An acceptably high accuracy of 70% supports an expectation that the oncogenes could distinguish carcinomas form normal tissue. As it could have been expected, response of the oncogenes was weak when dealing with benign tumours and normal tissue. However, unpaired t-test analysis found a few significant relations between the oncogenes and the lesion classes. In general, misclassifications could be explained with the data quality (noise) and with responsiveness of the oncogenes to certain type of tumour tissue.

Acknowledgements

Authors are grateful to Drs Prof. Marija Us-Krasovec, Ph.D., and Vera Kloboves-Prevodnik, M.Sc., for involving us in an interesting clinical discussion on diagnostic efficacy of the R-EGFR and Rcerb-B2 oncogenes which had initiated the present research.

References

[1] Willliam C. Dougali et. al., The neu-onkogene: signal transduction pathways, transformation mechanisms and evolving therapies
[2] Lori Jardines et. al., neu(c-erbB-2/HER2) and the Epidermal Growth Factor Receptor (EGFR) in Breast Cancer, Farhobiology 1993:61:268-252
[3] Giovanni Pauletti et. al., Detection and quantitation of HER-2/neu gene amplification in human breast cancer archival material using flourescence in situ hybridization, Oncogene 1996 13:63-72
[4] Pavešič N., Razpoznavanje vzorcev, in Slovene, ZAFER, 1992
[5] Duda R. O., Hart P. E., Pattern Classification and Scene Analysis, a Wiley Interscience Publication, 1973
[6] Niemann H., Pattern Analysis and Understanding, Springer-Verlag, 1981
[7] Gray J. W. Mastering Mathematica: Programming Methods and Applications, AP Professional, 1994
[8] Wolfram S., The Mathematica Book, Third Edition, Cambridge University Press, 1996