INFORMATICA, 2002, Vol. 13, No. 4, 455–464
© 2002 *Institute of Mathematics and Informatics, Vilnius*

Application of Data Mining Technique for Diagnosis of Posterior Uveal Melanoma

Darius JEGELEVIČIUS, Arūnas LUKOŠEVIČIUS
Institute of Biomedical Engineering, Kaunas University of Technology
Studentų 50–343, LT-3031 Kaunas, Lithuania
e-mail: arunas.lukosevicius@ktu.lt

Alvydas PAUNKSNIS, Valerijus BARZDŽIUKAS
Department of Ophthalmology, Institute for Biomedical Research
Kaunas University of Medicine
Eiveniu 4, LT-3007 Kaunas, Lithuania
e-mail: apaun@medi.lt

Received: August 2002

Abstract. Application of knowledge discovery in databases (data mining) for medical decision support is discussed in this work. The aim of the study was to use decision support algorithm for the differential diagnosis of intraocular tumors using parameters from eye images obtained by the ultrasound examination. Application of predictive modeling algorithm for decision tree formation using See5.0/C5.0 data mining system is presented. The decision tree was build using tumor geometry and microstructure parameters. The use of decision tree allows to differentiate tumors from other tumor-like formations. Low percentage of diagnostic errors shows that decision tree is reliable enough to offer it as "second opinion" for physician's decision support.

Key words: ophthalmic oncology, posterior uveal melanoma, ophthalmic ultrasound, data mining, decision tree.

Introduction

Use of information technologies in medicine enables acquisition and accumulation of diagnostic data in electronic form. The estimation of current clinical case in the form of a second opinion can be given for physician's decision support taking into account diagnostic parameters of all former clinical cases. Evaluation of such former cases can help physician in making clinical decision and diagnosis.

Data mining has become a popular technology in current research and for medical applications (Williams *et al.*, 1998; Kuo *et al.*, 2001). It is the extraction of implicit, previously unknown, and potentially useful information from data. The term 'Data Mining' refers to using a variety of techniques to process large amounts of information in order to discover knowledge, useful for decision-making. This covers a number of different approaches, such as clustering, data summarization, learning classification rules, finding dependency networks, analyzing changes, detecting anomalies, and so on. One of the successful applications of data mining techniques for medical diagnostic purposes is illustrated in (Kuo *et al.*, 2001). The diagnosis of breast tumor by data mining with effective use of decision trees is shown in this work. The See5.0 (C5.0) data mining tool was used here (Data Mining Tools See5 and C5.0.). The accuracy of the data mining for the breast tumor diagnosis was 96% comparing to the 86% of physicians diagnosis. This impressing result encourages searching for other applications and validating algorithm with actual data from ophthalmology.

One of the main examination methods of eye tumor is A/B ultrasound scanning. A/B ultrasound technique enables to localize, measure prominence and basis of the tumor and evaluate the shape, choroidal excavation, complications (retinal detachments, hemorrhages) and internal reflectivity.

Malignant uveal melanoma and metastasis belongs to the infrequent cancer diseases (only 0.2 percents of all human cancers) but they are the most frequent malignant eye tumors and represents one of the most lethal diseases seen in the ophthalmologic practice (Seregard, 1996). Early diagnosis and differentiation of the eye tumor is important for effective treatment of the patient.

Localization, dimensions, the shape and morphological type of the tumor are main parameters in diagnosis, prognosis and choosing treatment of the patient (Devron *et al.*, 1997). The evaluation of the morphological type of the tumor is possible only after the removing of the eye with the tumor or after diagnostic biopsy, but A/B ultrasound investigation enables to diagnose cell morphology of malignant intraocular tumors prospectively (Šebeliauskienė and Paunksnis, 2001). We used the See5.0 (C5.0) system for diagnosis of the eye uveal melanoma using diagnostic data of A/B ultrasound examination.

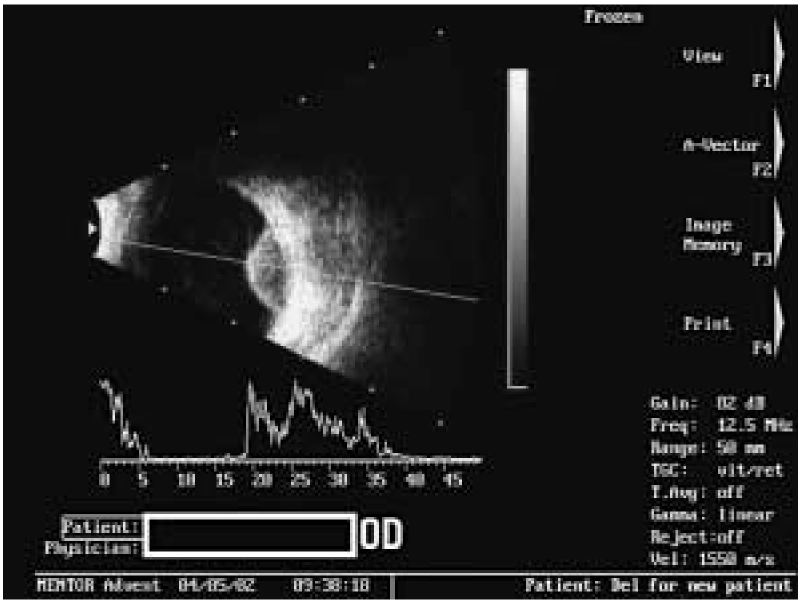
Purpose

The aim of the study was to test decision support algorithms for the differential diagnosis of intraocular tumors using parameters from eye images obtained by A/B ultrasound examination.

Materials and Methods

89 patients with intraocular pathology have been investigated in Eye Clinic of Kaunas University of Medicine. Intraocular tumors have been diagnosed for 46 patients, metastatic tumor of the eye – for 8 patients and were 35 clinically and echoscopically similar to the tumor cases (not tumors).

All patients have been examined with ultrasonic diagnostic imaging system *Mentor Advent* using A/B ultrasonic investigation mode, 12.5 MHz frequency, was set 80–90dB gain and 1550 m/sec velocity of ultrasonic wave (Fig. 1).



Gain: 82 dB
Freq: 12.5 MHz
Range: 50 mm
TGC: vit/ret
T.Avg: off
Gamma: linear
Reject:off
Vel: 1550 m/s

MENTOR Advent 04/05/02 09:38:18 Patient: Del for new patient

Fig. 1. Ultrasound B image of the tumor obtained by *Mentor Advent* A/B ultrasound system.

Diagnostic Parameters

Diagnostic parameters of tumor were calculated from A/B ultrasound images (Fig. 1). This procedure of parameterization is described in (Šebeliauskienė and Paunksnis, 2001; Mačiulis, 2002). Full list of parameters used for data mining in the diagnosis of intraocular tumors is presented in Table 1.

Table 1
Parameters of the case for data mining

Parameter	**Form of parameters values**
Sex	Verbal (Male/Female)
Tumors height	Continuous numeric, in mm
Tumors base	Continuous numeric, in mm
Shape	Verbal (Dome/Mushroom/Polygonal/Lobular)
Reflection coefficient	Continuous numeric, in %
Coefficient of regularity	Continuous numeric, relational
Choroidal excavation	Verbal (Yes, No, Unclear)
Retinal detachment	Verbal (Yes, No)
HB ratio (tumor height relation to tumor base)	Continuous numeric, relational

Data Mining Model

In data mining the greatest chance of success comes from combining expert's knowledge with advanced analysis techniques in which the computer itself identifies the relationships and features in the data (Williams *et al.*, 1998). The process of data mining creates models and/or rules from previously known data and diagnosis. After learning models are used for diagnostic prediction from current new data (Fig. 2). The techniques used to build these models are often referred to as machine learning or modeling. There are a number of machine learning and modeling technologies, including rule induction, neural networks, association rule discovery, clustering, and so on.

Decision Tree

Result of data mining can be presented in a convenient form of so called "decision tree", since it is presented in the form of a tree structure with decision rules. Decision trees are most commonly used for classification to predict what group a case belongs to. It can also be used for regression to predict a specific value. We chose the decision trees as the model in our study because the main purpose of our issue is to differentiate the tumors correctly.

Fig. 2 illustrates knowledge discovery system algorithm for diagnostic decision support. For the decision tree model data with diagnostic parameters and known diagnosis are used. The created decision tree is used for the new data processing and the result of it is certain diagnosis. This new data with known and approved diagnosis is used for decision tree model remodeling and improvement based on the new experience. When the ammount of data used for decision tree modeling (or in other words – for learning) gets sufficiently big the knowledge discovery system reaches its potentially best decision

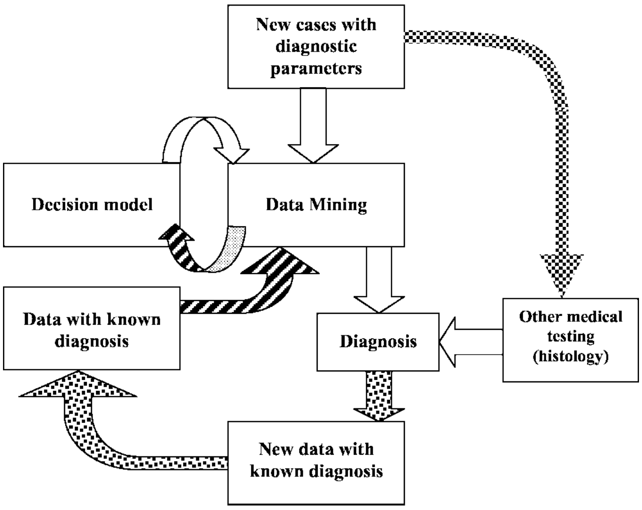

New cases with diagnostic parameters
Decision model
Data Mining
Data with known diagnosis
Diagnosis
Other medical testing (histology)
New data with known diagnosis

Fig. 2. Knowledge discovery system for medical diagnostic decision support.

support performance. The decision tree model shows importance of the diagnostic parameters. This importance can vary dependent to the data size and parameters should not be rejected even if it seems to be not very functional.

See5.0 (C5.0) is one of the most popular inductive learning tools originally proposed by J.R.Quinlan as C4.5 algorithm (Quinlan, 1993). The names of algorithms C4.5 and C5.0 are used as the names of data mining tools and software. Features and interrelations are ordered using a principle of maximization of expected information in these algorithms. Finally these algorithm constructs classification rules in form of a decision tree.

Technique for Decision Tree Synthesis

The used C5.0 algorithm is an extension of ID3 attribute-based machine learning system (Quinlan, 1993). The ID3 algorithm operation involves inducing classification rules for a given set of objects that are described by fixed set of attributes. The set of objects is known as a training set. It is assumed that it is known whether each object in the training set is a positive instance (P) or a negative instance (N). The system then applies an algorithm to induce a rule which would correctly classify the objects in the training set.

The ***basic ideas*** behind decision tree formation are that (Quinlan, 1993; Mitchell, 1997):

- In the decision tree each node corresponds to a non-categorical attribute and each arc to a possible value of that attribute. A leaf of the tree specifies the expected value of the categorical attribute for the records described by the path from the root to that leaf. This defines what a "Decision Tree" is.
- In the decision tree at each node should be associated the non-categorical attribute which is most informative among the attributes not yet considered in the path from the root. This establishes what a "Good" decision tree is.
- Entropy is used to measure how informative is a node. This defines what is meant by "Good".

Basic ideas are implemented by Decision Tree Learning *Algorithm* (ID3). Getting Examples, Target and Attributes as inputs, it works as follows from this simplified sequence:

- Create a root node
- If all Examples have the same Target value, give the root this label
- Else if Attributes is empty label the root according to the most common value
- Else begin
    - Calculate the information gain for each attribute, according to the average entropy formula
    - Select the attribute, A, with the lowest average entropy (highest information gain) and make this the attribute tested at the root
    - For each possible value, v, of this attribute
        - ◁ Add a new branch below the root, corresponding to $A = v$
        - ◁ Let Examples(v) be those examples with $A = v$

```
            ◁ If Examples($v$) is empty, make the new branch a leaf node labeled with
              the most common value among Examples
            ◁ Else let the new branch be the tree created by ID3(Examples($v$), Target,
              Attributes – $\{A\}$)
• end
```

C5.0 is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees and rule derivation.

Reduced-error pruning examines trees that have already been created by running a decision tree algorithm on a set of training instances, makes adjustments to the tree and then tests them on validation instances. The algorithm keeps changes only if the pruned tree performs better than the original tree.

To develop a training set, the algorithm can use part of the original instance set to create a decision tree and then use the rest as a validation set. The use of a validation set helps to eliminate over fitting data by dealing with inconsistencies and coincidences in the training set. Reduced-error pruning involves selecting internal nodes, re-labeling them with a classification and then discarding their descendent nodes. Rule post-pruning changes the paths to leaf-nodes along the decision tree to logical rules. Once these rules have been created, their antecedents are refined in an attempt to limit over fitting. The algorithm classifies instances by applying the rules in a sorted order (sorted by estimated accuracy).

C5.0 handles continuous-valued functions by dividing them into a set of discrete-valued functions. This can be repeated at each step of the algorithm to make the divisions that yield the largest information gain.

C5.0 can deal with missing attributes in several ways. One is to give the missing attribute the value that is most common for other instances at the same node. Or, the algorithm could make probabilistic calculations based on other instances to assign the value. C5.0 does not solely select which attribute to test on predicted information gain. Basing attribute selection on the gain ratio avoids selecting attributes that classify entities arbitrarily.

C5.0 supports boosting of decision trees. Boosting is a technique for generating and combining multiple classifiers to give improved predictive accuracy. By this process error rate is reduced on some datasets. C5.0 incorporates variable misclassification costs. Algorithm allows a separate cost to be defined for each predicted/actual class pair; if this option is used, C5.0 then constructs classifiers to minimize expected misclassification costs rather than error rates.

Automatic winnowing of attributes can be used in C5.0, discarding those that appear to be only marginally relevant before a classifier is constructed. Winnowing can lead to smaller classifiers and higher predictive accuracy.

Results and Discussions

89 patients with intraocular pathology have been investigated. Diagnostic parameters were calculated and prepared for the data mining with See5.0 (C5.0). The diagnoses of

intraocular findings was known. There were three types of diagnosis: tumor, metastatic tumor and not tumor (echoscopicaly similar to the tumor).

The See 5.0 system was used with its default settings, i.e., with no boosting and winnowing, due to the small data set amount.

At the first step there were 8 mistakes (9%) using decision tree build on this data. It means that in 8 cases from 89 decision tree was unable to set an a-priori known diagnoses.

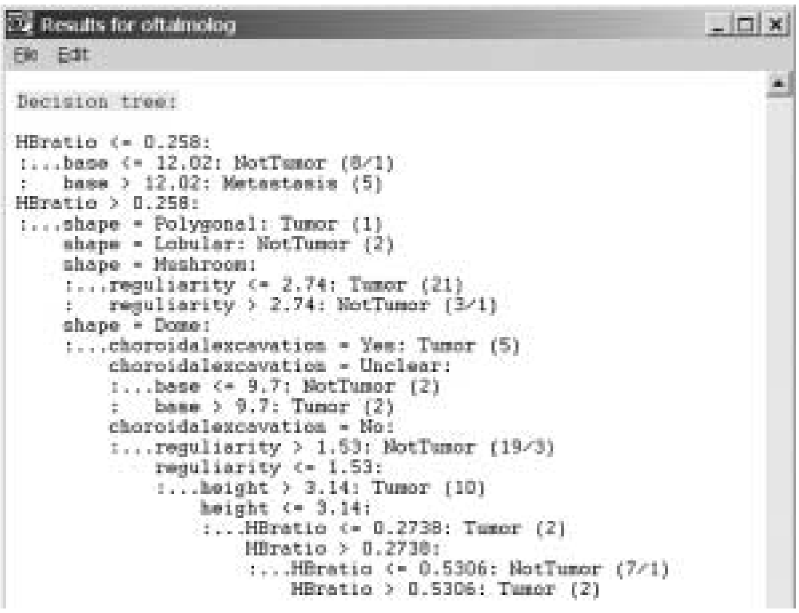
The See 5.0 system had recognized only 7 diagnostic parameters for decision tree synthesis. Rest 2 parameters – *sex* and *retinal detachment* data were unused. This is a positive distinctive feature of the algorithm, since there is no need to select voluntary most informative parameters for particular decision. Algorithm evaluates them and after pruning gives the verdict what parameters can be skipped. This saves time and efforts in acquisition of data. At the same time there are no limitations to include possibly informative parameters and to get an evaluation of it's usefulness for decision making.

After two parameters (*sex* and *retinal detachment*) were excluded from the data the next step was taken. In the new decision tree 6 parameters were taken (see Fig. 3). New composition of parameters led to decision tree where one more parameter (reflection coefficient) was skipped as not informative for tumor diagnostics.

After second step there were only 6 diagnostic errors (6.7%) using decision tree model presented on Fig. 3.

Despite of fully automatic decision tree synthesis, it is easily readable. This is another positive feature of this approach to compare with neural networks, logistic regression and other methods.

In the decision tree each node corresponds to the diagnostic parameter and each arc –



```
Decision tree:

HBratio <= 0.258:
:...base <= 12.02: NotTumor (8/1)
:   base > 12.02: Metastasis (5)
HBratio > 0.258:
:...shape = Polygonal: Tumor (1)
    shape = Lobular: NotTumor (2)
    shape = Mushroom:
    :...reguliarity <= 2.74: Tumor (21)
    :   reguliarity > 2.74: NotTumor (3/1)
    shape = Dome:
    :...choroidalexcavation = Yes: Tumor (5)
        choroidalexcavation = Unclear:
        :...base <= 9.7: NotTumor (2)
        :   base > 9.7: Tumor (2)
        choroidalexcavation = No:
        :...reguliarity > 1.53: NotTumor (19/3)
            reguliarity <= 1.53:
            :...height > 3.14: Tumor (10)
                height <= 3.14:
                :...HBratio <= 0.2738: Tumor (2)
                    HBratio > 0.2738:
                    :...HBratio <= 0.5306: NotTumor (7/1)
                        HBratio > 0.5306: Tumor (2)
```

Fig. 3. Decision tree made by the See5.0 data mining tool on intraocular ultrasound testing data.

to a possible value of that parameter. A leaf of the tree specifies the expected value of the diagnosis for the records described by the path from the root to that leaf. In parenthesis over diagnostic attributes are shown total number of cases in training set that satisfied the rules and the number of mistakes of that branch from the root.

The threshold and sequence of parameters for optimal classification of intraocular tumors into three groups in decision tree are presented. Dependent to decision tree in every one new case the prognosis and probability of possible diagnosis is given as an answer for physician. For example the diagnostic case with assumed tumor with height of 6.88 mm, base width 11.76 mm, mushroom shape, 1.22 regularity coefficient is evaluated as intraocular tumor with probability of 0.92.

Conclusions

The pilot decision tree for the differential diagnosis of intraocular tumors based on parameters from eye images obtained by A/B ultrasound examination, using the See 5.0 (C5.0) data mining tool was created. This model is intended to support physician's clinical decision as "second opinion". The main errors appeared in metastatic tumors classification (3 errors in 8 cases) because of small amount of cases.

Reliability of decision support will increase with the every increment of learning cases. Using the knowledge discovery scheme (Fig. 2) there is no need to interfere in the process or change something in programs. Only duty of physician is to feed the system with reliable data and diagnoses for consecutive learning of algorithm and use decision support in clinics. Same parameters also can be used to support other related clinical decisions as well.

References

Data Mining Tools See5 and C5.0. `http://www.rulequest.com/see5-info.html`

Devron, H.C., S. Kroll, T.L. Phillips (1997). Uveal melanoma. *Arch Ophthalmol*, **115**, 1014–1018.

Kuo, W.-J., R.-F. Chang, D.-R. Chen and Ch.Ch. Lee (2001). Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. *Breast Cancer Research and Treatment*, **66**, 51–57.

Mačiulis, A., A. Kopustinskas, A. Paunksnis and D. Šebeliauskienė (2002). Evaluation of echographic parameters for classification of intraocular tumors and retina degenerations. *Elektronika ir elektrotechnika*, **5** (40), 76–80 (in Lithuanian).

Mitchell, T. (1997). *Machine Learning*. McGraw Hill. `http://www-2.cs.cmu.edu/~tom/mlbook.html`

Seregard, S. (1996). Posterior uveal melanoma. *Acta Ophthalmol Scand*, **74**, 315–339.

Šebeliauskienė, D., and A. Paunksnis (2001). Echographic differentiation of malignant intraocular tumors. *Ultragarsas*, **41** (4), 25–28.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann. San Mateo, CA.

Williams, G., M. Hegland and S. Roberts (1998). *A Data Mining Tutorial*. Presented at the Second IASTED International Conference on Parallel and Distributed Computing and Networks (PDCN'98). `http://wwwmaths.anu.edu.au/~steve/pdcn.pdf`

D. Jegelevičius in 1995 graduated from Faculty of Telecomunications and Electronics, Kaunas University of Technology. In 1997 he received master degree in electronics engineering. On 2002 he defended his doctoral disertation on diagnostic signal procesing in telemedicine systems. Now he is working in Institute of Biomedical Engineering at Kaunas University of Technology.

A. Lukoševičius received a doctoral degree in electronics engineering from Kaunas Polytechnic Institute (KPI) in 1975 and subsequently joined the Ultrasonic Laboratory in the Department of Radioelectronics in KPI. In 1981 he became an assistant professor, and in 1994 was appointed a head of Research in the Laboratory of Ultrasonic Biomedical Engineering. In 1996 he received the degree of doctor habilitus and was appointed Vice-Rector for Research of Kaunas University of Technology. In 1998 he became a full professor. His research interests include signal processing, ultrasound interaction with non-homogeneous media, diagnostic systems and methods, biomedical engineering and telemedicine. He is a member of International Eye Research Society (ISER), the Society of Ultrasound in Ophthalmology (SIDUO), and is president of the Lithuanian Society of Biomedical Engineering, member of editorial boards of Lithuanian journals, coordinator of international research projects. He is a winner of Lithuanian National Science Award (2000).

A. Paunksnis is MD (habil.), professor, a head of Department of Ophthalmology of Institute for Biomedical Research of Kaunas University of Medicine. In 1969 graduated from Kaunas University of Medicine. Academic achievements: habilitated doctor since 1993, professor of ophthalmology since 1997. Areas of interest are: telemedicine, non-invasive ultrasound ophthalmologic investigation, ophthalmooncology.

V. Barzdžiukas is MD, associate professor in Department of Ophthalmology of Institute for Biomedical Research of Kaunas University of Medicine. In 1976 graduated from Kaunas University of Medicine. Doctoral disertation was defended on 1988. Areas of interest are: non-invasive investigation of ocular circulation (Doppler ultrasound, etc.), telemedicine.

Žinių analizės metodo taikymas akies gyslainės melanomos diagnostikai

Darius JEGELEVIČIUS, Arunas LUKOŠEVIČIUS, Alvydas PAUNKSNIS,
Valerijus BARZDŽIUKAS

Šiame darbe yra nagrinėjamas žinių analizės duomenų bazėse metodo panaudojimas diagnostinių medicininių sprendimų palaikymui. Darbo tikslas buvo panaudoti žinių analizę vidinių akies auglių diagnostikai, remiantis ultragarsinio tyrimo duomenimis. Naudotas prognozuojantis sprendimų medžio formavimo algoritmas (See5.0/C5.0 žinių analizės sistema). Panaudojant auglių geometriją ir mikrostruktūrą aprašančius parametrus, sudarytas sprendimų medis, leidžiantis atskirti auglius nuo kitų darinių. Mažas diagnostinių klaidų procentas rodo, kad toks medicininių sprendimų palaikymo būdas yra pakankamai patikimas "antrosios nuomonės" gydytojui šaltinis.