SELECTING THE MOST ACCURATE FORECASTING METHOD FOR MEDICAL DIAGNOSIS. BREAST CANCER DIAGNOSIS

A Case Study

Marc Almiñana, Alejandro Rabasa, Laureano Santamaría
Centro de Investigación Operativa, Universidad Miguel Hernández, Avda. Universidad s/n, Elche, Alicante, Spain

Laureano F. Escudero
Departamento de Estadística e I. Operativa, Universidad Rey Juan Carlos, c/ Tulipan s/n, Móstoles, Madrid, Spain

Antonio F. Compañ
Departmento Patología y Cirugía, Universidad Miguel Hernández, Crta. Nacional, N-332 s/n – San Juan, Alicante, Spain

Agustín Pérez-Martín
Departmento de Estudios Económicos y Financieros, Universidad Miguel Hernández, Elche, Alicante, Spain

Keywords: Data Mining, Forecasting methods, Diagnosis, Breast cancer.

Abstract: Different methods are usually applied for medical diagnosis problems. Most of them are only based on expert knowledge and the results are provided by model-driven methods and they are built from inflexible mathematical expressions. In this paper we suggest a Data-Driven perspective to facilitate the medical expert labour on diagnosis tasks. Furthermore, this paper offers a step by step procedure to select the most accurate forecasting method depending on the nature of the variables and the structure problem constraints. To validate such a selecting procedure, we apply it to a breast cancer diagnosis problem as a real case study.

1 INTRODUCTION

This paper addresses the state of the art in the different predictive methods used for clinical diagnosis, highlighting the advantages from a Data-Driven perspective, which dispenses with any physical or mathematical model that governs the problem domain and where a prediction is only based on historical data. The state of the art is summarised in a table, which classifies the different predictive methods in order to later design a procedure (based on them) which permits the selection of the most suitable method for each type of problem. Next, a diagnostic problem for breast cancer is proposed based on a public mammography data bank and a suitable method for the problem is determined following the proposed procedure and showing some of the results that are reached after its application. This paper concludes that using the procedure of predictive selection methods is convenient and it also gives details of the advantages of the rule systems classification for diagnostic prediction problems when dealing with unique variables of a non numerical nature.

2 STATE OF THE ART

2.1 Predictive Methods. A Data-Driven Perspective

There are very many different predictive methods, according to the type of variables capable of managing the objective to be reached and the domains they are applied in. The group of Data Mining methods which infer the behaviour of

Almiñana M., Rabasa A., Santamaría L., F. Escudero L., F. Compañ A. and Pérez-Martín A. (2010).
SELECTING THE MOST ACCURATE FORECASTING METHOD FOR MEDICAL DIAGNOSIS. BREAST CANCER DIAGNOSIS - A Case Study.
In *Proceedings of the Third International Conference on Health Informatics*, pages 142-148
DOI: 10.5220/0002590701420148

variables based only on historical data (methods belonging to the Data- Driven perspective) are presented as a magnificent alternative, (Solomatine 2002 [a] and 2002 [b]), in all those situations where the equations or models that govern the problem (Model-Driven methods) are unable to reflect the future state of the target variable accurately. From a very general perspective, Neural Networks, Decision Trees and Rule Systems derived from them, such as Genetic Algorithms, are considered the most representative of Data-Driven methods.

2.2 Predictive Methods in Medicine

Below, the main predictive methods used in the medical environment are outlined, indicating what type of problem is to be solved in each case. However, the nature of medical data in itself (as seen in section 3) along with the demand for an extremely high degree of accuracy makes it practically impossible to associate each problem or area of Medicine with an optimum predictive method. Besides generic expert systems (Lemke, Müller, 2001), (Suwa et al. 1982) there are many areas in Medicine where Data Mining predictive methods have been applied, from the prediction of diabetic disorders, (Mugambia et al. 2004) and digestive disorders, (Gorzalczany, Gradzki, 1999) to diagnoses in ophthalmology, (Shi et al. 2006); cardiovascular diseases (Gamberger et al. 2002) and (Podgorelec et al. 2005) or haemodialysis treatments, (Kusiak et al. 2004), including the main medical objective: detection and prevention of cancer, where early diagnoses have a vital role. There are numerous studies that use Data Mining techniques for analysing data for prostate cancer, (Tahir, Bouridane, 2006); cervical cancer, (Ho et al. 2004), and especially breast cancer, (Polat et al. 2005) and (Kohli et al. 2006).

(Park et al. 2006) give a very complete and rigorous summary of different logistic regression applications: Decision Trees, Neural Networks and Case Based Reasoning (Nilsson, Sollenborn, 2004), covering papers published from 1993 to present day. In the said study, Neural Networks are shown to still be widely used in medical diagnosis and the most used Neural Network in this Framework is the multi-layer perceptron.

Neural Networks are usually used (also in Medicine) in combination with other techniques, mainly: Decision Trees, among which we can highlight CHAID, CART, C4.5 (and its variation C5.0) for their good performance in predictive problems.

Regression systems have been chosen as a classical predictive method on numerous occasions (Kurgan, Cios, 2004). This is the same for Neural Networks which have gradually incorporated Diffuse Logic in order to adapt to problems where numerical thresholds should be smoothed out, (Gorzalczany, Gradzki, 1999). Likewise, given the nature of the data (chaotic in many cases, and almost always with incomplete values) there are numerous medical studies that are based on the Rough Set Theory, (Wang et al. 2006); (Kusiak et al. 2004) and (Pattaraintakorn et al. 2005).

It is worth mentioning the wide variety of studies that use Decision Trees as a predictive method for medical problems. The algorithm C4.5 has been successfully applied on numerous occasions: (Chan et al. 2006); (Tahir, Bouridane, 2006) and (Polat et al. 2005). It is also frequent to find Decision Tree applications that are improved with the use of Genetic Algorithms in order to optimize the generated Rule Set, (Podgorelec et al. 2005), and even in combination with the abovementioned Rough Set theory, (Kusiak et al. 2004), in order to generate Decision Trees based on data with a lot of inconsistencies. There are different comparative studies about the use of Decision Tees, which put C4.5 and CHAID as the highest accuracy ratio, (Block et al. 2006); (Guler, Gurgen, 2004) and (Ho et al. 2004). With respect to the different measures used, besides the prediction ratio, in order to quantify the adaptation of Rule Systems generated with medical data, some authors choose flexible concepts according to sensitivity and specificity, (Mol et al. 1999) and (Timm, 1998). Other papers are based on different variations to the measures derived from the Bayes Theorem, (Shortliffe, Buchanan, 1975) and (Kukar, Groselj, 2005). The majority of these comparative studies are carried out on different data bases from the UCI repository (Machine Learning Repository, University of California), as with the proposed case study.

In Medicine, on many occasions, the Rule Systems generated are so extensive that it becomes necessary to apply a reduction method and even generate confirmation rules which are oriented so as to be contrasted by professionals from the domain (Gamberger, 2002) who help to simplify the final rule system, so that it is more legible and easier for the experts to interpret. In (Almiñana et al. 2008) we propose the reduction of rules for diagnosing thyroid disorders. Table 1 shows a summary of the methods for solving predictive problems in the field of medicine.

Table 1: Summary of predictive methods in the medical area.

THEORY STUDIES
(Block et al. 2006). Forecast methods Comparison. C4.5
(Guler, Gurgen, 2004). Forecast methods Comparison
(Ho et al. 2004). CHAID. Cervical cancer.
(Nilsson, Sollenborn, 2004). CBR: Case Based Reassoning
(Park et al. 2006). CBR: Case Based Reassoning
(Timm, 1998). Sensibility and specificity

CLASSICAL METHODS
(Kohli et al. 2006). 0-1 Integer Program. Rules, breast cancer
(Kurgan, Cios, 2004). Logistic Regression
(Shortliffe, Buchanan, 1975). Bayes Theorem

TOOLS
(Lemke, Müller, 2001). Knowledge Miner
(Shi, 2006). LASSO. Patterns in Ophtalmology
(Suwa et al. 1982). Expert Systems

ROUGH SET
(Pattaraintakorn et al. 2005). Attribute selection
(Wang et al. 2006). Cancer forecasting

DECISION TREES (DT) AND RULE SYSTEMS (RS)
(Chan et al. 2006). C4.5 forecasting posology
(Gamberger et al. 2002). Confirmation rules. Cardio
(Kusiak et al. 2004). DT + Rough Set
(Mol et al. 1999). Flexible. vs no-flexibles RS
(Mugambia et al. 2004). Diabetes and trauma
(Podgorelec et al. 2005). Genetic Algorithm for optimal rules
(Polat et al. 2005). C4.5 variable reduction, breast cancer
(Tahir, Bouridane, 2006). C4.5 and RR-TS algorithm. Prostata cancer

3 THE NATURE OF PREDICTIVE PROBLEMS IN MEDICINE

From the references studied, it is possible to draw up a series of characteristics which are common to the majority of the medical problems where prediction is required. Firstly, the data bases are completed by the analytical records of patients (on occasions from different data origins) where there may be numerous absent values which on occasions need to be completed in the data preprocessing stage. Besides, on several occasions, these values are of a numerical nature, while experts may need to handle discreet values, therefore, the discretization process of values is critical. When the data bases contain data collected through patient monitoring systems in real time, the information is usually already ordered and ready to be processed. Another characteristic common to the majority of the predictive problems in Medicine is the high level of accuracy required. Finally, when Diagnostic Help Systems are being dealt with, discriminating rules are especially useful as certain clinical symptoms can be ruled out from their antecedents.

4 SELECTION OF THE MOST APPROPRIATE PREDICITIVE METHOD

The most commonly used predictive methods in Medicine have been quoted. However, what are the circumstances that make one method more appropriate than another? And even more importantly, is there a procedure for choosing the most suitable method for each case?

In fact, the predictive method is not associated to the clinical speciality which it is to be applied to, but depends on:

- the continuous or nominal nature of the variable to be predicted
- the nature and characteristics of the data it is based on
- the type of predictive model which is to be obtained

If the objective is to predict the value of a nominal value (discreet), the problem corresponds to a Regression task. If on the other hand, the target variable is numerical (continuous) it is a Classification task.

Table 2: Methods, tasks and algorithms.

METHOD → ***TASK ↓***	**NEURAL NETWORKS**	**REGRESSION SYSTEMS**		**GENETIC ALGOR.**	**STOCHASTIC NETWORKS**	**DECISION TREES**	
		LINEAR/ NON-LINEAR	**LOGIST.**			**for CLASSIF.**	**for REGRESS.**
CLASSIFIC.	Backprop.		Logist. Regress.	G.A.	Bayessian Networks	ID3 C4.5	
REGRESS.	Backprop.	Linear/ Non-Linear		G.A.			C4.5

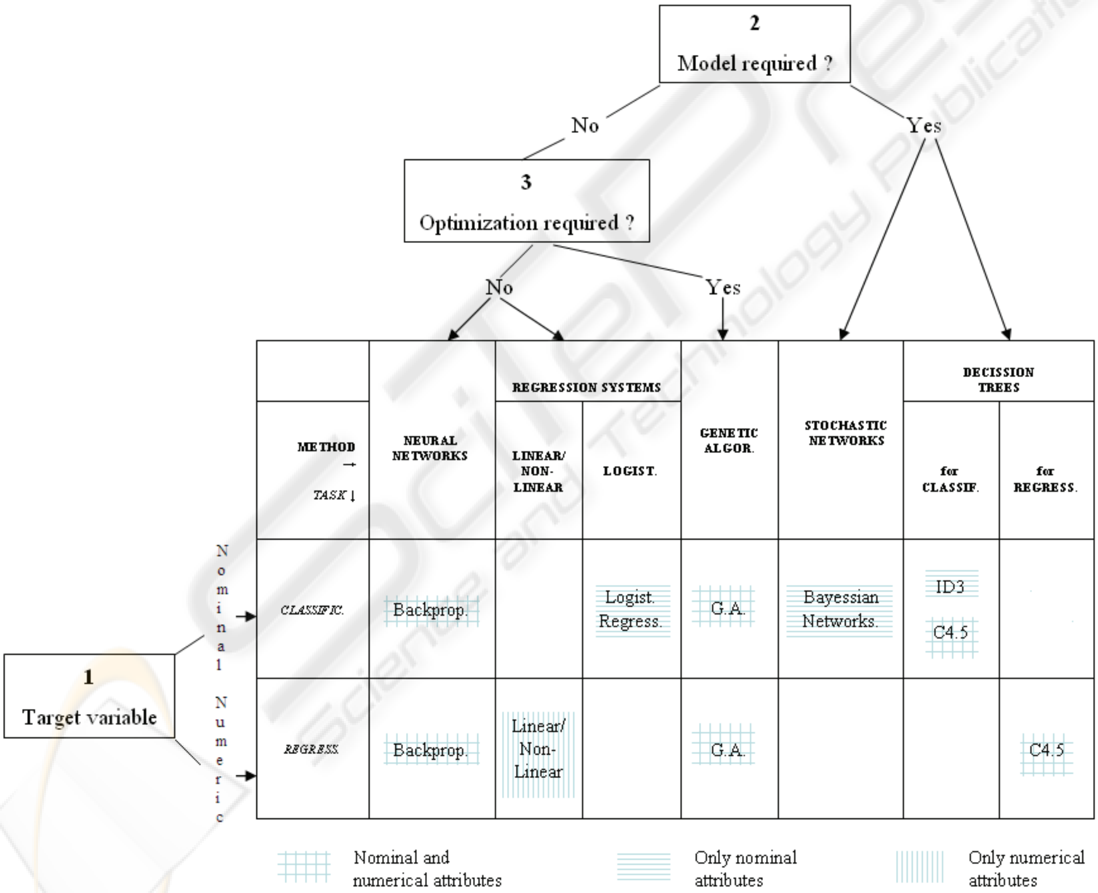

Figure 1: Procedure to select the most accurate forecasting method.

Table 2 gives a summary of the different predictive methods divided into their respective tasks: Classification and Regression. Each square contains the name of the algorithm or the most significant generic procedure of the group, which implements the corresponding method.

This classification should not be understood in a rigid way, but as a classification guide. On many occasions, as stated in section 2.2, these algorithms are applied in combination.

Table 3: Attributes and values for mammography data base.

Attribute	Description	Possible values
BI-RADS	A priori Risk Evaluation	1, 2, 3, 4, 5 (from minimum to maximum risk) ? (missing value)
AGE_D	Patient's age	<40 (<40 years), 40-50 (40 ... 50 years), >60 (>60 years),
SHAPE	Mass form	1 (round), 2 (oval), 3 (lobular), 4 (irregular), ? (Missing value)
MARGIN	Mass margin	1 (circumscribed), 2 (microlobulated.), 3 (obscured), 4 (ill-defined), 5 (speculated), ? (Missing value)
DENSITY	Mass density	1 (high), 2 (iso), 3 (low), 4 (fat-containing), ? (Missing value)
SEVERITY Consequent	Diagnosis	0 (benign), 1 (malign)

The procedure proposed for this study, in order to choose the most adequate predictive method, is the following:

Step 1: what type of target variable is it?

- Nominal: the focus is centred on the row "Classification" Tasks

- Numerical: the focus is centred on the row "Regression" Tasks

Step 2: Is the generation of a model required at the end of the process?

- No: Step 3

-Yes: the focus is centred on the rows "Stochastic Networks" and "Decision Trees"

Step 3: Is optimization required?

- Yes: the focus is centred on the column "Genetic Algor."

- No: the focus is centred on the columns "Neural Networks" and "Regression Systems"

Step 4: The rows and columns on which the focus has been centred cross. In the case where more than one method is obtained, then its ability to manage the type of attributes (nominal or numerical) of the problem to be solved is taken into account.

Fig.1 shows a diagram of the procedure described above.

5 CLASSIFICATION RULES FOR BREAST CANCER DIAGNOSIS, A CASE STUDY

The problem of predicting breast cancer is proposed as a case study, using rule systems that indicate to the expert the probability of benign or malign cancer, based on the values of the antecedent variables. To do so, the public database from the University of California, Irvine, which consists of 961 mammographies, is used. The variables registered are shown in Table 3.

The application of the proposed procedure for the choice of the predictive method to be applied would be as follows:

Step 1: Target variable, severity, is nominal (benign, malign). The focus is established on the row Classification Tasks.

Step 2: The generation of the predictive model of rule systems is required. The focus is centred on the columns Stochastic Networks" and "Decision Trees". To be exact, the rule systems are generated through the in depth path of the Decision Trees.

Step 3: Not involved

Step 4: On crossing the row and the column, the focus is centred on the method "Decision Trees for Classification" which can be carried out through algorithms from the ID3 and C4.5 family. Because all of the variables are nominal or previously discretized, the algorithm ID3 appears as the ideal candidate for solving this predictive problem.

After applying the corresponding decision tree, some of the more interesting classification rules that are generated based on the in depth path of the tree are:

- The variable which in itself best serves for the diagnosis is BI-RADS.

Example: If BI-RADS=5 (A priori maximum risk), then SEVERITY=1 (malign). Confidence=88,4%

- If along with this variable the DENSITY variable is considered, the average correlation with the diagnosis improves notably.
Example: If BI-RADS=5 and DENSITY=3 (low), then SEVERITY=1 (malign). Confidence=89,9%
- The second variable that, considered along with BIRADS, obtains high degrees of average correlation with the diagnosis is SHAPE.
Example: If BI-RADS=5 and SHAPE=4 (irregular), then SEVERITY=1. This rule having Confidence=90,8%.
Besides, elimination rules are also apparent, which are especially useful in clinical diagnosis.
Example: If BI-RADS=4 (A priori high risk) and SHAPE=1 or 2 (round or lobular), then SEVERITY=0 (benign). Confidence=90,7%, 90,2% respectively

In this case study, the joint consideration of more variables does not lead to more accurate diagnoses.

6 CONCLUSIONS

From the points presented in this paper, it can be concluded that the most appropriate method does not depend on the target medical speciality of the study but on the real target of the prediction, the nature of the data which are involved and the need (or not) to obtain a predictive model at the end of the process.

Although in practice the combination of two or more methods is very frequent, the step by step execution of the proposed procedure for the selection of the most suitable method leads to only one optimum predictive method.

In the face of nominal and univariable clinical diagnostic problems (for example SEVERITY= benign or malign), the classification rules that are derived from the in depth route of the ID3 type decision trees, appear as a very reliable predictive method which is easy for experts to interpret. Besides, this type of predictive model highlights the combination of optimum variables and their degree of correlation with the diagnosis, permitting the design of more reduced analyses, which can allow for more reduced analysis times, less invasive or even more economical procedures

REFERENCES

Almiñana, M., Escudero, L.F., Pérez, A., Rabasa, A., Sánchez, C., Santamaría, L., 2008. Reducting Classification Rule Systems Applied To Thyroid Functional Diagnosis. *Proceedings XXIV International Biometric Conference*. University College Dublin

Block, P., Paern, J., Hüllermeier, E., Sanschagrin, P., Sotriffer, C., Klebe, G. , 2006. Physicochemical Descriptors To Discriminate Protein–Protein Interactions In Permanent And Transient Complexes Selected By Means Of Machine Learning Algorithms. Wiley Inter Science. *Proteins: Structure, Function, and Bioinformatics* 65, 607–622

Chan, A.L., Chen, J.X., Wang, H.Y. , 2006. Application Of Data Mining To Predict The Dosage Of Vancomycin As An Outcome Variable In A Teaching Hospital Population. Dustri-Verlag. *International Journal of Clinical Pharmacology and Therapeutics* 44 , 11, 533-538

Gamberger, D., Lavrac, N., Krstacic, G. , 2002. Confirmation Rule Induction And Its Applications To Coronary Heart Disease Diagnosis And Risk Group Discovery. IOS Press. *Journal of Intelligent and Fuzzy Systems* 12 , 1, 35-48

Gorzalczany, M.B., Gradzki, P. , 1999. Computational Intelligence In Medical Decision Support -A Comparison Of Two Neuro-Fuzzy Systems. *Proc. ISIE'99*. Bled, Slovenia

Guler, N., Gurgen, F.S. , 2004. The Effects Of Data Properties On Local, Piecewise, Global, Mixture Of Experts, And Boundary-Optimized Classifiers For Medical Decision Making. Springer-Verlag. *Computer and Information Sciences, Proc. Lecture Notes in Computer Science* 3280, 51-61

Ho, S.H., Jee, S.H., Lee, J.E., Park, J.S. , 2004. On Risk Factors For Cervical Cancer Using Induction Technique. Elsevier. *Expert Systems with Applications* 27, 97–105

Kohli, R., Krishnamurti, R., Jedidi, K. , 2006. Subset-Conjunctive Rules For Breast Cancer Diagnosis. Elsevier. *Discrete Applied Mathematics* 154, 1100 – 1112

Kukar, M., Groselj, C. , 2005. Transductive Machine Learning For Reliable Medical Diagnostics. Springer Science+Business Media, Inc. *Journal of Medical Systems* 29, 1

Kurgan, L.A., Cios, K.J. , 2004. CAIM Discretization Algorithm. IEEE Computer Soc. *IEEE Transactions* *on Knowledge and Data Engineering* 16, , 2, 145-153

Kusiak, A., Dixon, B., Shah, S. , 2004. Predicting Survival Time For Kidney Dialysis Patients:A Data Mining Approach. Elsevier. *Computers in Biology and Medicine*. , accepted

Lemke, F., Müller, J-A. , 2001. Medical Data Analysis Using Self-Organizing Data Mining Technologies. *Systems Analysis Modelling Simulation* 43 ,10, 1399-1408

Mol, B.W., van der Veen, F., Bossuyt, P.M.M. , 1999. Implementation Of Probabilistic Decision Rules Improves The Predictive Values Of Algorithms In The Diagnostic Management Of Ectopic Pregnancy. European Society of Human Reproduction and Embryology. *Human Reproduction* , 14 11, 2855-2862

Mugambia, E.M., Hunterb, A., Oatleyd, G., Kennedy, L. , 2004. Polynomial-Fuzzy Decission Treestructures For Classifying Medical Data. Elsevier. *Knowledge Based Systems*, 17, 81-87

Nilsson, M., Sollenborn, M. , 2004. Advancements And Trends In Medical Case-Based Reasoning: An Overview Of Systems And System Development. American Association for Artificial Intelligence. *Malardalen University*. Technical Report

Park, Y.J., Kim, B-Ch., Chun, S-H. , 2006. New Knowledge Extraction Technique Using Probability For Case-Based Reasoning: Application To Medical Diagnosis. Blackwell Publishing Ltd. *Expert Systems*, 23

Pattaraintakorn, P., Cercone, N., Naruedomkul, K. , 2005. Hybrid Intelligent Systems: Selecting Attributes For Soft-Computing Analysis. *Proceedings of the 29th Annual International Computer Software and Applications Conference*

Podgorelec, V., Kokol, P., Stiglic, M.M., Hericko, M., Rozman, I. , 2005. Knowledge Discovery With Classification Rules In A Cardiovascular Dataset. Elsevier. *Computer Methods and Programs in Biomedicine* 80 Suppl. 1, 39-49

Polat, K., Sahan, S., Kodaz, H., Gunes, S. , 2005. A New Classification Method For Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System , FS-AIRS. Springer-Verlag. *Advances in Natural Computation 2, Proc. Lecture Notes in Computer Science* 3611, 830-838

Shi, W., Wahba, G., Wright, S., Lee, K., Klein, R., Klein, B. , 2006. LASSO-Pattern Search Algorithm With Application To Ophthalmology Data. *University of Wisconsin*. Technical Report, no. 1131

Shortliffe, E.H., Buchanan, B.G. , 1975. A Model Of Inexact Reasoning In Medicine. *Mathematical Biosciences* 23, 351-379

Solomatine, D.P. , 2002 [a]. Data-Driven Modelling: Paradigm, Methods, Experiences. *Proc. 5th International Conference on Hydroinformatics*, Cardiff, UK.

Solomatine, D.P. , 2002 [b]. Applications Of Data-Driven Modeling And Machine Learning In Control Of Water Resources. Idea Group Publishing. *Computational Intelligence in Control*, 197-217

Suwa, M., Scott, A.C., Shortliffe, E.H. , 1982. Completeness And Consistency In A Rule-Based System. AAAI. *The AI Magazine* 3, 16-21

Tahir, M.A., Bouridane, A. , 2006. Novel Round-Robin Tabu Search Algorithm For Prostate Cancer Classification And Diagnosis Using Multispectral Imagery. IEEE-Inst Electrical Electronics Eng. *IEEE Transactions on Information Technology in Biomedicine* 10 , 4, 782-793

Timm, I.J. , 1998. Automatic Generation Of Risk Classification For Decision Support In Critical Care. Ed. Bellazzi and Zupan, Brighton, *UK. ECAI '98 Workshop Notes on 'Intelligent Data Analysis in Medicine and Pharmacology*

University of California, Irvine. UCI Standard Repository, http://archive.ics.uci.edu/ml/

Wang, X., Yang, J., Jensen, R., Liu, X. , 2006. Rough Set Feature Selection And Rule Induction For Prediction Of Malignancy Degree In Brain Glioma. Elsevier. *Computer methods and programs in biomedicine* 8, 3, 147–156